South African Radio League

Introduction

to

Amateur Radio

A study guide for the
Radio Amateur Examination

Edition 1.2
January 2016

ISBN 978-0-620-69471-1

Table of Contents

Version Control

Version	Date	Description	Contributors
1.0	2005	Original version of the "*Radio Amateur Examination Manual*"	Author Andrew Roos ZS1AN
1.1	2009	Addition of digital systems	Wessel du Preez ZS5BLY Peter Hers ZS6PHD Colin de Villiers ZS6COL Ean Retief ZS1PR George Honiball ZS6NE Mark Zank ZS6YES Maarten du Preez ZS6ZY Mickey Esterhuyzen ZS5QB Rassie Erasmus ZS1YT
1.2	2016	Renamed "Introduction to Amateur Radio" Added preface Editorial revision and formatting Rearranged chapters Added QSL cards and LotW Added field-effect transistors Added voltage sources and models Added description of sine function Added frequency ranges Added multi-pole filters and ringing Added practical components at RF Added light-emitting diodes Added battery capacity and recharging Added field-effect transistors Added thermionic tubes Added integrated circuits Added WSJT Added error correction Added receiver thermal noise Added Pi and T networks Added reciprocal mixing Added log-periodic arrays Added feeders, including waveguides Added VSWR Added antenna stacking and phasing Added multi-path and fading Added auroral scatter and backscatter Added propagation prediction and link budgets Added direct radiation and shielding Removed ASCII table Added convolution Added safety sections: human body, lightning Added description and use of callsigns Added emergency operations Added international regulations and IARU band plans Added glossary of abbreviations Added section numbers Added UTC and log keeping suggestions Added typographic conventions for greek letters	Chris R. Burger ZS6EZ With valuable input from: Vincent Harrison ZS6BTY Colin de Villiers ZS6COL

The SARL gratefully acknowledges the American Radio Relay League's permission to use diagrams in Chapter 30. Most of these diagrams were extracted from the ARRL book *Experimental Methods in RF Design* by Wes Hayward W7ZOI, Rick Campbell KK7B and Bob Larkin W7PUA.

This version: 1.2.21 2016-03-07

SARL, Box 1721, Strubensvallei, 1735 South Africa www.sarl.org.za

Preface

This *Introduction to Amateur Radio* is exactly that: The document provides an entry into the wonderful world of amateur radio, specifically for would-be radio amateurs.

South Africa complies with CEPT T/R 61-01, the European recommendation on mutual recognition of amateur radio licences. South African radio amateurs can exercise their privileges in more than 80 countries and territories with minimal paperwork. In exchange, we have an obligation to ensure that all licence holders possess a minimum level of knowledge commensurate with that required in other countries.

Andrew Roos, then ZS1AN, produced most of this study guide around 2005. Colin de Villiers ZS6COL brought it into alignment with ISO units and formatting in 2007. It covers all items in the HAREC syllabus, which forms Annex 6 to TR 61-02. The latest version of that syllabus is dated 2004-02-12. It is through Andrew's and Colin's hard work and with mediation from the South African Radio League that this monumental piece of work is offered free of charge to anyone wanting to become a radio amateur in South Africa. Our hope is that its availability will also help would-be amateurs elsewhere in Africa. Amateur radio has a great role to play in fostering skills in modern communications technology. There is no better way to learn than to tinker with things in practice!

We have refrained from including lots of pictures, specifically to keep the document size manageable for downloading. You will find plenty of images on the Internet to supplement the descriptions.

You may notice a few rough edges in the document. Remember that it is a volunteer effort, and an ongoing project. We had to let this version go so that it can be used, while we continue to polish some more. Please contribute your insights and observations to the Editor, so that the next edition will be even better.

Enjoy the process of learning about amateur radio. Many of us were shaped by amateur radio. It provided exposure to the world outside South Africa when the Internet did not exist, and when our passports were not universally welcome. It offered hands-on technical exposure that opened doors to technical innovation and careers, making South Africa a world leader in many fields of communications.

But above all, it will change your life. You can form lifelong friendships with people you haven't met. You can visit long-time friends when you travel the world. You can learn new languages. You can even string a wire over a tree and talk to the world, with equipment that comfortably fits into the palm of your hand.

Arthur C. Clarke's Third Law states:

> *Any sufficiently advanced technology is indistinguishable from magic.*

He was right, you know. It really is all magic.

Chris R. Burger ZS6EZ
CSIR Meraka Institute
Editor: Edition 1.2
Pretoria, January 2016

Chapter 1: Overview of Amateur Radio

Amateur radio is a hobby that involves experimenting with radio (and related technologies like television or radar) for fun and education. It is also known as “Ham Radio” and radio amateurs are sometimes referred to as “hams”. Like most hobbies, there are many different activities that fall under its umbrella.

1.1 Communicating With Other Amateurs

Using radio to communicate with other amateurs is one of the foundations of the hobby. Most amateurs have a radio station of their own, but radio clubs often establish stations for communal use. Increasingly, shared stations can be accessed via the Internet so that participants can access the station from any convenient location.

Stations could range from very simple to very elaborate. A combination of a *transmitter* and a *receiver* is known as a *transceiver*. Handheld transceivers are available for talking to others in the same town, while small transceivers suitable for portable or mobile use can achieve world-wide communication. Antennas make a huge difference to the results, and can range from a simple arrangement of wires to huge antenna farms featuring several tall towers.

Radio amateurs communicate in many different *modes*. The term refers to the way information is encoded onto the radio signal. Radio signals started with Morse code telegraphy, where the radio signal is simply switched on and off in accordance with a standard code. This mode is known as Continuous Wave (or CW), and is still popular in amateur radio after more than a century. Phone is a collective term for any speech-based modes, including FM, AM and SSB, where a normal human voice is transmitted via radio and interpreted by ear. Finally, many digital modes allow data to be transmitted via radio. These include FSK, PSK, WSPR, WSJT, SSTV and many others. These modes allow the transmission of text and pictures, much like the Internet does.

Don’t worry too much about these terms; their meaning will become clear later.

Amateur radio contacts (also known as *QSOs*) range from the briefest exchange of callsigns lasting mere seconds to long conversations (known as *rag-chews*) that may last hours.

Amateur radio is not like the phone system since you generally can’t contact a particular station on demand. If you want to speak to a particular person, you must agree to a time and a frequency where you will meet. Such an arrangement is known as a schedule, or “sked” for short. Even then, there is no guarantee that propagation will allow the stations to hear one another.

Otherwise you can just speak to whoever happens to be listening and is interested in a chat, which is a great way to meet new friends and learn new things. There are also some regularly scheduled networks (or “nets”) where operators who share a common interest get together at a particular time and frequency to exchange ideas.

1.2 Collecting QSL Cards

After communicating with another amateur station, it is customary to send a confirmation in the form of a QSL card. This card is a postcard with information about yourself and your station, and details of the QSO such as the date, time, frequency, mode and the callsign of the station worked. Many amateurs take a great deal of pride in their QSL cards, which are often works of art. As well as being something to display and a nice reminder of the contact, QSL cards are required if you wish to claim a contact for an award (see below).

At the end of the twentieth century, several electronic replacements for QSL cards started evolving. The most important of these, the *Logbook of the World,* confirms the correctness of contacts, but does not include facilities for pictures. As a result, many other systems have sprung up, with *eQSL.cc* perhaps being the most popular.

1.3 Building Radio and Electronics Equipment

Many amateurs build at least some of their equipment. Some build equipment from purchased kits or from plans found in amateur radio magazines or on the Internet. Others build their equipment from scratch, doing all the necessary design and sourcing the components themselves. The complexity ranges from simple projects, such as a computer soundcard interface that can be built in an evening, to complete radio transmitters and receivers that may take months or years of work. Today, microprocessors and digital signal processing (DSP) is an increasingly important part of the hobby, so building equipment may also involve writing the necessary micro-controller or DSP programs.

Of course, for those who do not enjoy electronics, everything you need to participate in the hobby can be purchased off the shelf.

1.4 Building Antennas

Many amateurs find the complexity of modern transceivers to be beyond their construction capabilities. However, antennas provide a ready target for experimentation. Simple antennas can be made with string and wire, and more elaborate installations can include heavy hardware and advanced automation.

Most amateurs build at least some of their own antennas. Antenna projects can be very rewarding as good results may be obtained from fairly simple designs. A number of software packages allow you to design an antenna and model its performance before you invest in the construction of the antenna.

1.5 Public Service and Emergency Communications

Radio amateurs have a proud history of making their skills and equipment available for public service and emergency communications. On the public service side, amateurs provide communications for many sporting events such as rallies, marathons and cycle tours where their ability to communicate effectively from remote places is of great assistance to the organisers.

Many amateurs also ensure that their radio stations have some alternative power source, such as batteries, a generator, or solar power, so that they can continue to provide communications in the event that the telephone and power distribution systems are disrupted by a natural or manmade disaster. In South Africa, Hamnet, a special interest group of the South African Radio League, coordinates amateur emergency communications.

1.6 DXing

“DXing” means communicating with as many different places as possible, often in order to qualify for certificates and awards. The term comes from the use of “DX” as a telegraphy abbreviation for “long distance”.

There are many different awards, mostly in the form of a handsome certificate. Some of the harder awards also include a trophy or plaque. Major awards include:

- The premiere award is the *DX Century Club (DXCC)*, which requires proven communications with at least 100 different countries. Over 60 000 amateurs have earned DXCC, including more than 200 South Africans.

- *Worked All ZS*, for contacting 100 stations in the various regions of South Africa. The award's name comes from the fact that "ZS" is the most common prefix assigned to South African amateur radio stations. Almost 600 of these certificates have been issued, mostly to South Africans.

- The *Islands on the Air (IOTA)* award series, for communicating with stations located on islands.

- *Worked all States (WAS)*, given for contacting all 50 states of the USA.

- *Summits on the Air (SOTA)*, for communicating with mountaintop stations.

- *Worked All Zones (WAZ)*, for proven contacts with all 40 CQ Zones. Some zones are easy, while others are sparsely inhabited or propagationally challenging.

1.7 DXpeditions

Because DXers are always on the lookout for countries, islands, mountains or provinces that they have not worked before, there is often a flurry of interest and activity when a rare country or island is activated by some intrepid radio amateur. Expeditions to unusual places for the purpose of setting up and operating a radio station there are called "DXpeditions", and participating in DXpeditions is itself a very rewarding and challenging activity.

1.8 Contests

Contests bring out the competitive nature of some radio amateurs, who enjoy the challenge to contact as many different stations as possible over a predetermined period of anything from an hour or two up to 48 hours or more. Contests may be run on a local, national, regional or international basis and may attract anything from 10 to 5000 contestants. Many contests have several entry categories to allow similarly equipped stations to compete with each other.

1.9 Satellite Communications

The amateur community has successfully launched a number of small communications satellites for the use of radio amateurs around the world. Communicating with other amateurs via satellite (or via the earth's natural satellite, the moon) gives radio amateurs an unparalleled opportunity to learn about the technology that underlies much of the modern era of communications. Because amateurs themselves develop these satellites as a cooperative, non-profit venture, those who are interested in the design and construction of satellites also have the opportunity to study the designs and may eventually be able to contribute to new amateur satellite projects.

1.10 Maritime and Off-Road Communications

The maritime and off-road communities are increasingly turning to amateur radio for their communication needs. Thousands of small craft such as yachts make use of the services provided by maritime nets which pass on weather reports and crucial safety information, allow mariners to access email and assist in the search for missing boats. Off-roaders who venture into uninhabited areas can also benefit from amateur communications, both between vehicles within a party and also back to a "home base" or to summon assistance in an emergency.

1.11 Licence Requirements in South Africa

In order to operate an amateur radio station, you must have a licence issued by the Independent Communications Authority of South Africa (ICASA). When you are issued with a licence, you will also be given a unique callsign. Every amateur station has a callsign, which is used to identify it on the air. Regulations require every transmission to be identified.

South African amateur callsigns consist of the letters “ZR”, “ZS”, “ZT” or “ZU” followed by a single digit indicating the region of the country in which you are located, followed by one to three letters. For example, the original author of this document had a callsign “ZS1AN”. The “ZS” indicates an Unrestricted licence, the “1” shows a location in the Western Cape, and the letters “AN” are a unique identifier of that particular station.

Amateur radio callsigns are a source of great pride, and become closely associated with the individual. It is not uncommon to refer to individual amateurs solely by their callsigns!

There are two different licence classes:

The Class A Licence (ZR or ZS callsign)
This licence has full privileges on all frequency bands.

To obtain an Unrestricted (ZR or ZS) licence you must pass the full Radio Amateur’s Examination (Class A). You must also demonstrate the ability to install and operate amateur radio equipment intended for the High Frequency (HF) bands. This manual provides the study material for the Class A examination., so if you pass the examination at the end of the course you will be entitled to a Class A licence.

The Class A licence is also recognised internationally under the CEPT agreement, as a Class 1 licence. CEPT is a European organisation that harmonises telecommunications regulation in the European Union, but several other countries are also signatories. Under CEPT, a South African Class A licence holder can operate from any country that supports CEPT with very simple paperwork. Countries covered include most of the European Union, the USA, Canada, Israel, Australia and New Zealand. The syllabus for this study guide is based on the HAREC syllabus which underlies the CEPT agreement.

The Class B Licence (ZU callsign)
The Class B licence is an entry-level licence that provides youngsters with a simple entry point into the hobby. It has restricted privileges in the High Frequency (HF) and Very High Frequency (VHF) bands, mostly with a maximum transmitter power output of 100 W. for Single Sideband (SSB) transmissions. To obtain a Class B licence, you must pass a simplified Radio Amateurs’ Examination. Class B licences are only issued up to the age of 20, and lapse at the holder’s 25th birthday. Holders are encouraged to upgrade to a Class A licence before the expiry date.

1.12 The Radio Amateurs’ Examination

The Radio Amateur’s Examination is held twice each year, in May and October. It consists of two papers: *Regulations and Operating Procedures* and *Technical*. The *Regulations and Operating Procedures* paper has 30 multiple-choice questions and the *Technical* paper consists of 60 multiple-choice questions. In order to pass the examination you must obtain a minimum of 50% on each of the papers and an overall mark of at least 65%. If you pass one paper but fail the other, you can rewrite just the paper you failed at the next sitting of the examination, provided that you pass sufficiently well to get 65% overall.

The examination is set and administered by the South African Radio League (SARL), the national organisation representing radio amateurs in South Africa. The Independent Communications Authority of South Africa (ICASA) is a statutory body that regulates the communications industry. The examination fee changes from time to time, so ask your course instructor what the current fee is, or consult the SARL Web site, http://sarl.org.za.

1.13 Restrictions on the Use of an Amateur Radio Station

The Radio Regulations include some restrictions on the use of an amateur radio station. It is important that you understand these in case you find that what you had planned to do with your amateur radio licence is not permitted!

1. Amateur radio stations may not be used for broadcasting. Amateur radio is intended for direct “one-on-one” communications with other amateurs, and not as a community broadcasting service.

2. Amateur radio stations may only transmit music under very specific conditions, which are intended to ensure that they do not become pirate broadcast stations.

3. No products or services may be advertised on amateur radio.

4. Amateur radio stations may not transmit messages for reward.

5. Amateur radio stations may not be used to transmit business messages that could be sent using the public telecommunications service.

6. Amateur radio stations may not be used to transmit indecent, offensive, obscene, threatening or racist comments.

7. Amateur radio stations may not be used to pass third-party traffic (in other words, messages that originate from anyone other than the amateur who is operating the station) except during an emergency.

This chapter briefly outlines what amateur radio is all about, what the licence requirements are, and what legal restrictions there are on what can be transmitted by amateur radio stations. I hope you have decided that amateur radio is a hobby that you wish to participate in. Welcome you to the amateur community! We hope that you will find this course interesting and worth your while.

Chapter 2: Operating Procedures

2.1 International Regulations

The ITU

The International Telecommunications Union (ITU) is an agency of the United Nations, tasked with regulating international telecommunications. It is headquartered in Geneva, Switzerland. Its building houses an amateur radio station, callsign 4U1ITU, which can be heard on the air regularly and which counts as a separate country for amateur radio purposes.

The ITU publishes a set of international radio regulations, which govern all countries and all services. Our South African regulator, the Independent Communications Authority of South Africa (ICASA) is tasked with implementing and enforcing regulations to manage all radio services within South Africa, in compliance with ITU regulations. The latest version of the ITU Radio Regulations, dated 2012, can be downloaded from their Web site, itu.int.

The amateur radio service is defined by the ITU as follows:

> *A radiocommunication service for the purpose of self-training, intercommunication and technical investigations carried out by amateurs, that is, by duly authorized persons interested in radio technique solely with a personal aim and without pecuniary interest.*

An amateur radio station is simply defined as a station in the amateur service.

You will notice that self-training is one of the prime purposes of the amateur service. We are all here to learn!

Article 25 of the ITU Radio Regulations governs amateur radio. There are only 13 rules, spanning less than two pages. National regulators are left a lot of flexibility in the details of how amateur radio is regulated. The amateur satellite service is also included by reference in Article 25.

The ITU defines three regions of the world, within which frequency allocations and other collective rules are fairly homogeneous. Africa falls into Region 1, along with Europe.

CEPT

CEPT is the French abbreviation for the European Conference of Postal and Telecommunications Administrations (cept.org). It serves a purpose similar to that of ITU, but applicable only to European countries.

The reason why CEPT is important to us is that South Africa complies with the requirements of T/R 61-01, a Recommendation of CEPT that governs international licencing of CEPT amateurs. This manual complies with the Harmonised Amateur Radio Examination Certificate (HAREC) syllabus. As a result, any amateur holding a licence issued by a national authority that has implemented T/R 61-01 can operate in the territory of any other such country. In our case, we can operate almost anywhere in Europe, the USA, Canada, Israel, Australia and New Zealand without special paperwork. This concession also applies to possessions of all these countries, such as US islands in the Pacific, most Dutch and French islands in the Caribbean and even French possessions in South America and the Pacific. CEPT T/R 61-01 saves huge amounts in licence fees and reams of paperwork when travelling abroad.

If you want to operate in a foreign country using your South African callsign, download T/R 61-01 from the Web. As this paragraph is being written, the latest version is dated

2015-01-05. You must carry certain documents with you at all times (including your South African licence, your passport and your HAREC). The HAREC can be obtained from the SARL if you are an SARL member. You must use a callsign as prescribed in the table. For example, let's look at the South African callsign ZS1AN. More information about callsigns follows later in this chapter. For the moment, let's accept that this callsign is recognisable as a South African callsign from the Western Cape. The holder has to sign ZA/ZS1AN in Albania, OE/ZS1AN in Austria and EW/ZS1AN in Belarus. The USA has more than a dozen compulsory prefixes, depending on location. In Texas, W5/ZS1AN would be appropriate, while in Hawaii, KH6/ZS1AN would be correct.

The IARU

The International Amateur Radio Union (IARU) (iaru.org) represents the interests of amateur radio world wide. It has 155 member societies covering most territories. In South Africa, the member society is the South African Radio League (SARL).

Apart from advocacy at the ITU's rulemaking World Radio Conferences, the IARU also establishes voluntary band plans for amateur use. These band plans normally extend across an ITU region. South Africa would therefore adhere to Region 1 band plans.

In some cases, band plans are included in national regulations, making compliance compulsory. In other cases, band plans are simple gentlemen's agreements, and compliance is purely voluntary. In general, do not operate contrary to a band plan. These band plans are carefully formulated to take all interests of amateurs into account, and represent a compromise that hurts everyone equally. The considerations that went into making the band plan may not be obvious to you, and you may cause unanticipated damage if you decide to act contrary to the band plan.

Callsigns

With few exceptions, each licenced radio station in the world has a callsign. The purpose of the callsign is to identify transmissions from that station for the purpose of mitigating interference and to facilitate regulation.

An amateur radio callsign consists of a number of characters (letters and digits), normally between three and six. Examples of callsigns include W1AW, 7P8Z, 3DA0Z and ZS1AN.

Non-amateur callsigns have different formats. In South Africa, callsigns such as ZRB, ZS-RSA and ZRAM8100 are used in other services, and do not conform to the format for amateur radio callsigns.

Each amateur radio callsign consists of a prefix and a suffix. The suffix normally consists of one, two or three letters. The prefix contains at least one letter and one digit. The exact rules are beyond the scope of this text, but the prefixes above are W1, 7P8, 3DA0 and ZS1 respectively. The suffixes are AW, Z, Z and ASF.

The prefix normally indicates the area in which the station operates, while the suffix indicates the individual station. In the examples above, W1 is the northeastern part of the USA known as New England. 7P is Lesotho. 3DA indicates Swaziland. ZS1 indicates the Western Cape province of South Africa. W1AW is the official station of the American Radio Relay League, a callsign inherited from its founder, Hiram Percy Maxim.

Callsigns can also contain appendages, to indicate special conditions or locations that differ from the place of issue. W1/7P8Z would indicate that the Lesotho station 7P8Z is operating in New England. 7P8Z/P indicates that the station is portable, not at its normally licenced location, but inside Lesotho. 7P8Z/M indicates that the station is mobile, somewhere in Lesotho. ZS1AN/5 indicates that the station is operating somewhere in KwaZulu Natal

(ZS5). 7P8Z/MM is maritime mobile, somewhere on the world's oceans, on a Lesotho-registered vessel[1].

There are also special callsigns that do not comply with these rules. A recent example is ZS90SARL, a special callsign used by the SARL to commemorate its ninetieth anniversary. This callsign has a double digit in the prefix and a four-letter suffix, clearly not in accordance with normal rules. Similar callsigns are allocated from time to time, mostly for special events.

By law, callsigns must be used regularly in each contact. In some countries, the callsign of the station itself and the station it is addressing must be used in each transmission. Also in some countries, displaying the callsign on the equipment is compulsory. Either way, ensure that you do not make transmissions without prominent and intelligible use of your callsign. Practical guidelines are provided below.

2.2 HF Phone Procedures

HF means "high frequency", also known as Shortwave. On these bands, world-wide communication is possible using simple antennas. Local contacts are more often made on VHF (Very High Frequency) bands using repeaters. Suitable procedures for those conditions are discussed later. This section deals with Phone (voice) contacts on the HF bands.

The Phonetic Alphabet

The phonetic alphabet is used whenever information must be spelt out. It should be used for callsigns when initiating a contact. This alphabet was painstakingly optimised for international communications, taking into account differences in the way people hear and speak sounds. Although some amateurs may think it is cute to use their own phonetics, doing so hampers intelligibility and places an unnecessary burden on someone whose mother tongue may not resemble English at all.

Once it is clear that the other station has got your callsign correct then you can revert to normal pronunciation ("ZS1AN" instead of the phonetic "Zulu Sierra One Alpha November"). Although the prefix is fairly self-explanatory in local communications, the suffix should always be spelled out if there is any possibility of confusion. In this example, the difference between "N" and "M" is difficult to hear, so the suffix should always be given phonetically.

Alpha	*Hotel*	*Oscar*	*Victor*
Bravo	*India*	*Papa*	*Whiskey*
Charlie	*Juliett*	*Quebec*	*X-ray*
Delta	*Kilo*	*Romeo*	*Yankee*
Echo	*Lima*	*Sierra*	*Zulu*
Foxtrot	*Mike*	*Tango*	
Golf	*November*	*Uniform*	

Note that some of these pronunciations are a little unexpected. Many operators get J, P and Q wrong.

Knowing the words included in the phonetic alphabet aids intelligibility in poor conditions. For example, if you hear someone with a heavy accent saying something that sounds like "Bof", you know the word must have been "Golf".

[1] Huh?

Initiating Contacts
Before calling, you should listen for at least 30 s to see whether the frequency is clear. If you do not hear anyone else on or near the frequency, you can ask whether the frequency is clear:

> *Is this frequency in use?*

You may not be able to hear the station speaking at the moment, but there may be someone else who can hear both you and the station speaking, who will then respond that the frequency is in use.

Wait another few seconds. If you have not heard anything then you can proceed to call "CQ" to ask for a contact.

> *CQ CQ CQ, this is Zulu Sierra One Alpha November, Zulu Sierra One Alpha November, Over.*

Wait for at least 5 s. If you have not received a response, you can call again. If after you have tried many times you still have not received a response, the lack of response may indicate that propagation conditions are poor on the band you have chosen, or that you need more antenna work...

If you want, you can make a *directional* call, which means asking for only certain stations to reply. If you call *CQ DX*, you are asking for only "long distance" (DX) contacts, which usually means stations on another continent. There are exceptions. On VHF or UHF, a station from 500 km away would be considered DX, as normally only line-of-sight contacts are expected. If you call "CQ Europe" or "CQ Germany", you are asking only for stations from a particular continent or country to reply.

Responding to a CQ
If you hear a station calling CQ and you would like to make contact, check the following before you call:

1. Is it a directional call, and if so are you in the right area to respond? For example, a South African station should not respond to "CQ Japan" but may respond to "CQ Africa" or to "CQ DX" from a non-African station.

2. Make sure you know where the station is listening for a response. Most stations will listen on their own frequency. However, rare DX stations may work "split" which means they are listening on a different frequency, generally higher than the one they are calling on. For example, if you hear a DX station call "Sierra Tango Zero Romeo Yankee up five" it means the operator will be listening 5 kHz higher than the frequency he transmitted on. You will need to know how to activate the "split" function on your transceiver to work this station, so that you can continue to listen on his transmit frequency while transmitting on his listening frequency.

3. Ensure that a suitable antenna is connected and (if necessary) that your antenna tuning unit (ATU) is correctly set for the frequency and antenna. If the ATU is not set, *do not* tune up on the frequency where you hear the CQ call, as you will cause interference to the station calling. Rather change frequency by at least 3 kHz to an unoccupied frequency, and after checking that the frequency is not in use, tune up there and then return to the frequency where you heard the call.

Of course it is wise to check these things *before* you search for stations calling CQ, so when you hear one you can respond immediately. Suppose you hear W1XX calling and having checked everything you are ready to call. Then you would say:

> *Whisky One X-ray X-ray, this is Zulu Sierra One Alpha November, Zulu Sierra One Alpha November, over.*

Note that the callsign of the station being called is always given *first*, and the callsign of the station calling comes *last*. This order is important and getting it wrong will mark you as a novice.

It is unnecessary to repeat the callsign of the station you are calling. Most people know their own callsigns! If you are uncertain of the station's callsign, ask specifically. As for your own callsign: If conditions are favourable and you have a powerful station, once should be enough. Good DX operators can make over 300 contacts an hour. Most callers only call once. Under poor conditions, however, you might want to repeat your callsign once or twice.

Exchanging Reports

After making contact, the first things stations do is usually to exchange signal reports and basic information such as the name and location of the operator. If the signal report indicates that the other station is not hearing you well, don't try to engage in an extended conversation. It will only frustrate you and the other operator when you cannot make yourself understood.

Signal reports are exchanged according to the standard Readability-Strength-Tone code (RST). The Tone part is only used for carrier-based communication (CW and RTTY), so for Phone it is RS—Readability and Strength only. The meaning of the RST code is as shown below:

Readability

1 -- Unreadable
2 -- Barely readable, occasional words distinguishable
3 -- Readable with considerable difficulty
4 -- Readable with practically no difficulty
5 -- Perfectly readable

Signal Strength

1 -- Faint signals, barely perceptible
2 -- Very weak signals
3 -- Weak signals
4 -- Fair signals
5 -- Fairly good signals
6 -- Good signals
7 -- Moderately strong signals
8 -- Strong signals
9 -- Extremely strong signals

Tone

1 -- Sixty cycle AC or less, very rough and broad
2 -- Very rough AC, very harsh and broad
3 -- Rough AC tone, rectified but not filtered
4 -- Rough note,some trace of filtering
5 -- Filtered rectified AC but strongly ripple-modulated
6 -- Filtered tone, definite trace of ripple modulation
7 -- Near pure tone, trace of ripple modulation
8 -- Near perfect tone, slight trace of modulation

9 -- Perfect tone, no trace of ripple or modulation of any kind

The S component is often associated with S meter readings. Some receivers have an S meter to measure the strength of the incoming signal. However, the S in RST is not directly related to the S meter reading. As you can see, S is actually a subjective assessment, not directly related to any measurement. However, the S meter reading can provide some guidance on which S report would be appropriate.

In the early days of radio, T was often lower than T9. With modern transceivers, almost all signals are T9. If not, remedial action is required.

Readability is strongly dependent on signal-to-noise ratio, so it is perfectly possible to get reports like 519 (faint signals but perfectly readable) or 399 (extremely strong but readable with considerable difficulty).

To simplify paperwork, DXpedition and contest stations that make many thousands of contacts often hand out 59 or 599 reports to everyone. If you want to make their task easier, you might want to do the same. Do not be surprised to get a 59 or 599 report even after the station has repeatedly struggled to copy your callsign…

The Q code "QTH" is often used to mean the location of the operator, although phone operation allows the use of normal language, especially between native English speakers. If you are working someone with limited English, Q codes are very useful.

You might hear the following reply from W1XX:

> *Zulu Sierra One Alpha November this is Whisky One X-ray X-ray. Thanks for the call, you are five and six, fifty-six. My name is Bob, Bravo Oscar Bravo, and my QTH is Boston, Massachusetts. ZS1AN from W1XX.*

The signal report indicates that our signal is perfectly readable, with reasonably good signal strength. You would reply with a signal report, and also your name and location:

> *W1XX from ZS1 Alpha November. Good morning, Bob, and thanks for the report. Your signal is five nine, five nine here in Cape Town. My name is Andrew, Alpha November Delta Romeo Echo Whiskey, Andrew. I'm testing a new rig here, a Kenwood TS850S, running 100 watts into a triband Yagi at 15 metres.. W1XX this is ZS1AN. Over.*

Under good conditions, you may want to omit the "over", as it should be obvious when your transmission is over.

And so the conversation continues. You must by law identify your station on each separate transmission (each "over").

Ending the QSO

"QSO" is also from the Q code, and it means a contact between two stations, which may include several transmissions ("overs") by each station. The end of the conversation will probably go something like this:

> *W1XX from ZS1AN. Well, Bob, thanks for the nice chat, I must be off now. I will QSL via the bureau. Greetings to you and your family and see you later. W1XX this is ZS1AN. Bye!*

> *ZS1AN from W1XX. Fine, Andrew. It was nice to meet you. Enjoy that new radio, it sure sounds good from here. All the best until next time from W1XX, clear.*

There are many cutesie phrases like “73”, “88” and “standing by” which some people use. They are strictly optional. There is no reason why you have to deviate from normal speech.

“73” is from an old telegraphy code meaning “best wishes”. It is already plural, so you should *not* say “73s” like many do! An alternative when addressing someone of the opposite gender is “88”, which means “love and kisses”. “QSL via the bureau” means send a QSL postcard confirming the QSO via the QSL bureau. The QSL bureau is a bulk delivery service for QSL cards that operates via the amateur radio societies in many countries, including the South African Radio League in South Africa.

After the Contact
If you have not already filled in your log during the QSO, you should do so immediately afterwards. Remember that you are required by law to keep a log of all HF transmissions (including unanswered CQs).

If you have offered to send a QSL card, it is a good idea to write it out immediately, as it can become a chore if you wait for hundreds of QSOs to accumulate before writing out the cards.

What Not to Do
Don’t put out endless streams of “CQ” calls. It is most irritating to hear

> *CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ This is Zulu Sierra One Alpha November CQ CQ CQ…*

There is very little information in the “CQ”. Emphasise your callsign, as that is what people want to hear. Don’t call CQ for longer than about 15 s before taking a listen. Few people are prepared to wait for several minutes before you listen, and will simply tune further down the band. Nothing prevents you from calling again if there has been no response.

Never ever say “good buddy” or “10-4” or any other CB jargon, or cute phrases like “over and out” (whatever that means) or “standing by”.

2.3 Telegraphy Procedures

Telegraphy generally takes place through Continuous Wave (CW) transmissions. You will later learn exactly what this term means, but for the moment, just accept that CW means Morse code telegraphy. Although Morse is no longer in use in commercial telegraphy or radio, it is alive and well in amateur radio.

CW QSOs generally follow a similar format to HF phone. The key to the enjoyment of CW is proficiency. There is only one way: Once you have learned the code, get your feet wet and practice until you become comfortable. Some of the most stimulating conversations on the ham bands are found on CW!

Pay some attention to sending equipment. These days, the easiest way to get on CW is to use your keyboard as a sending device, but you may consider getting a decent paddle and keyer. The paddle uses a sideways movement with your thumb and index finger to produce dits and dahs with precise timing. You can send good-sounding code at high speed with little effort. Trying to use a hand key is laborious and will reduce your enjoyment considerably. You will probably abandon CW completely and miss out on a fun aspect of ham radio.

Abbreviations
Most CW operators use a lot of abbreviations. If everything is spelled out at the typical speed of most amateur Morse operators, very little would get said. With the use of abbreviations, proficient Morse operators converse almost as quickly as on Phone.

There are two main kinds of abbreviations used: the Q Code, which uses three-letter groups starting with the letter Q to represent questions and answers; and informal abbreviations for commonly used words. Let's start with the Q Code. Each entry can be used either as a question—in which case it is followed by a question mark—or as a statement, which may be in response to the question. For example, "QTH?" means "what is your location?" and the reply might be "QTH Cape Town" meaning "my location is Cape Town". You should get to know the following abbreviations:

Code	Question	Statement
QRG	What is my exact frequency?	Your exact frequency is... [in kHz]
QRK	What is the readability of my signals?	The readability of your signals is [R]
QRL	Are you busy?	I am busy.
QRM	Are you being interfered with?	I am being interfered with.
QRN	Are you troubled by static?	I am troubled by static
QRO	Should I increase power?	Increase power.
QRP	Shall I decrease power?	Decrease power.
QRQ	Shall I send faster?	Send faster. [words/ minute]
QRS	Shall I send more slowly?	Send more slowly. [words/ minute]
QRT	Shall I stop sending?	Stop sending.
QRU	Have you anything for me?	I have nothing for you.
QRV	Are you ready?	I am ready.
QRX	When will you call me again?	I will call again at .... [UTC][kHz]
QRZ	Who is calling me?	You are being called by ....[callsign]
QSB	Are my signals fading?	Your signals are fading.
QSL	Can you acknowledge receipt?	I acknowledge receipt.
QSO	Can you communicate with … directly?	I can contact … directly. [callsign]
QSP	Will you relay (a message) to ...? [callsign]	I will relay to ... [callsign]
QSY	Shall I change frequency to ...? [kHz]	Change frequency to ... [kHz]
QTH	What is your location?	My location is ...

The comments in square brackets indicate what has to be sent, and should not actually be transmitted. For example, you might say "QRX 1730", or in common usage "QRX 5", which means that the other station must wait for your next transmission in five minutes.

Note that QRX also has the informal meaning "standby" and QRT means "shut down the station". The difference between QRM and QRN is that QRM means "man-made interference", while QRN means "natural noise". A "QRP" station means a station transmitting with low power, usually 5 W output or less.

Initiating Contacts
First listen for at least 30 s to see whether the frequency is in use. If nothing is heard, ask whether the frequency is use by sending:

QRL? de ZS1AN

"QRL?" means "are you busy?" or "is this frequency in use?". "de" means "from" and is used immediately before the callsign of the station transmitting the message. If you hear any response, find another frequency. The station hearing you is supposed to respond with a "QRL" or "C" (for "Si", or "Yes" in most Latin-based languages), meaning "yes, this frequency is busy, please go away". However, you will hear all kinds of responses, such as

"Y", or even "YES". If you hear any of these, find yourself another frequency and try again.

If no-one responds, you can call CQ, which is similar to the procedure on phone:

CQ CQ de ZS1AN ZS1AN K

Once again "de" identifies the callsign of the sending station. The single letter "K" at the end is an invitation to *any* station to reply, the equivalent to "over". Note that the phonetic alphabet is never used in Morse.

Send a little slower than you can comfortably receive. Most novices[2] can send faster than they can reliably copy, and set a trap for themselves by calling too fast. If you are a beginner, the person answering your call is probably comfortable at a much faster speed than you are, and will probably send slightly faster than you are sending. You may embarrass yourself when you cannot read what the other station is sending.

Replying to a CQ
As with phone, send the callsign of the station you are calling *first*, and your callsign *last*. For example,

W1XX de ZS1AN ZS1AN $\overline{KN}$

The "KN" at the end with the bar over it means "send the letters K and N together without leaving the normal space between letters". Since K is dah-di-dah and N is dah-dit, this symbol is dah-di-dah-dah-dit, which is an invitation only to the called station to reply. These symbols made up of two letters run together are known as *procedure symbols*.

Since the "$\overline{KN}$" is actually the Morse symbol for an open-bracket "(", we'll use that symbol in following transmissions.

The station will proceed to give you a signal report, usually along with his or her name and QTH.

ZS1AN de W1XX GE OM Tnx fer call ur RST 439 439 Name Bob Bob
QTH Boston MA Boston MA Hw? $\overline{AR}$ ZS1AN DE W1XX (

As you can see, lots of informal abbreviations are used:

GE	- good evening
OM	- old man, used to refer to any male operator
Tnx	- thanks
fer	- for, because it is quicker in Morse!
ur	- your
RST	- RST signal report
Hw?	- How did you receive this?

Of course, in "*Boston MA*", the MA is the ZIP code abbreviation for the state of Massachusetts.

$\overline{AR}$ (the bar indicates that it is a single symbol, di-dah-di-dah-dit) is a procedure symbol that means "end of message". You might reply

[2] And old timers with personality issues…

W1XX de ZS1AN R GM Bob Tnx fer rprt ur 56n 56n Name Andrew Andrew
QTH Cape Town Cape Town = Rig 100W to 3el Yagi =
Wx fine 25C 25C = OK? AR W1XX de ZS1AN (

Again a few new abbreviations:

R - Roger (meaning "I received everything correctly")
rprt - report
56n - "9" is often abbreviated to "n". The most common report is "5nn"
3el - three element
Wx - weather
25C - temperature 25 degrees celcius
OK? - did you receive this transmission OK?

Note that the single "R" sent at the start of the message means "I received everything you sent correctly". It is not necessary to spell this out; and conversely, you should not send "R" if you did not receive *everything*. The "=" sign stands for the "break" symbol dah-di-di-di-dah that is usually used to separate thoughts or sentences. You can fruitfully use this symbol as a time killer while composing your thoughts during the transmission.

Quick turnaround
Suppose you missed Bob's name on the first over. You could either send everything laboriously and include a request for him to say his name again, or you can quickly send:

W1XX de ZS1AN Pse name agn? BK

He would then send:

BK Bob Bob BK

after which you can continue with your transmission. "Pse name agn?" means "please send your name again?", while the "BK" is equivalent to the Phone term "Break". It can also be used when trying to break into an existing QSO (which you would obviously only do if you really had something to contribute, like if one of the QSO participants is your long-lost brother).

Other useful abbreviations include *msg* (message), *tx* (transmitter, transmit or transmission), *rx* (receiver, receive or reception) and *rcvd* (received).

Ending the QSO
You can send the Q code "QRU" ("I have nothing further for you") to indicate politely that you have run out of things to say and would like to end the QSO. Conversely, if a station sends QRU, that is not an invitation to tell him your life story, but an indication that he or she wants to finish the QSO. So it might go like this:

ZS1AN de W1XX R Tnx fer info es nice QSO = QSL sure via buro =
73 to u es urs hpe cuagn = QRU AR ZS1AN de W1XX (

A few more abbreviations:

es - and (it's shorter and faster in Morse)
urs - "yours", so "u es urs" means "you and yours"
hpe - hope, or I hope
cuagn - see you again, so "hpe cuagn" means "I hope to see you again"

Then we finish with:

> *W1XX de ZS1AN R Tks Bob QSL OK via buro es lotw 73 es cul my friend*
> *$\overline{VA}$ W1XX de ZS1AN TU*

"Tks" is another alternate abbreviation for "Thanks", along with "tnx" and "tu". CW operators really are a polite bunch!

LotW is the Logbook of the World, the ARRL's online system that allows instant confirmation of contacts for some awards. Once both you and W1XX have uploaded particulars of this QSO, both of you will obtain an electronic QSL that can be used to prove that you really did make contact on this date, time, frequency and mode. Most operators appreciate both LotW and paper confirmations.

"cul" means "see you later" and the procedure symbol $\overline{VA}$ (also written as $\overline{SK}$) means "end of QSO". The "TU" at the end is a final "thank you" and is often followed by two Morse dits (or "e e") as a final flourish.

2.4 Repeater Procedures

Repeater (n): *A device for increasing the range of poorly-equipped stations and reducing the range of well-equipped stations.*

Repeaters are used for local FM communications. They allow stations that might not have "line of sight" propagation to each other to still make contact as the repeater will relay the signal between the stations, as long as both have line of sight to the repeater. Repeaters are sited on high terrain to provide good coverage. In principle, any two stations in the coverage area can communicate. Repeaters are particularly useful for mobile or pedestrian stations with low power and small antennas. Well-equipped fixed stations can often get better coverage by themselves.

Some repeaters are also linked to other repeaters world-wide via the Internet Radio Linking Project (IRLP). In South Africa, there are two linked networks. The largest one is the Cape Linked Repeater Network, linking more than a dozen repeaters from Cape Town to Bloemfontein. Such repeaters really do extend your range, as you can talk to other mobiles halfway around the world.

The repeater consists of a receiver and a transmitter which retransmits everything that the receiver receives. All the 2 m repeaters in South Africa use a separation of 600 kHz between the input and output frequencies, with the input frequency being 600 kHz below the output frequency. The user would listen on the output frequency and transmit on the input frequency.

The repeater is activated ("keyed") when it receives a signal on its input frequency. This signal is then simultaneously retransmitted on the output frequency. When referring to the frequency of a repeater it is standard practice to refer to the repeater *output* frequency. The "145,750 MHz repeater" transmits on 145,750 MHz and receives 600 kHz lower, at at 145,150 MHz.

"kHz" is pronounced kilohertz and "MHz" is megahertz. There are 1000 kHz in 1 MHz. You will learn the exact meaning of these terms later. For the moment, let's just accept that these frequencies are in the 2 m amateur band.

Some repeaters are triggered falsely by noise or abusive users, much to the chagrin of listeners. The repeater my trigger every few minutes, causing a hiss until it turns itself off again. Even worse, it may retransmit the noise on the input frequency continuously, driving users up the wall. To circumvent this problem, many repeaters now have a CTCSS tone

squelch, which requires a continuous sub-audible tone of 88,5 Hz on your transmission before they are activated. You may need to set your radio to transmit a CTCSS tone to get into your local repeater.

QSOs on the repeater are much less formal than HF phone QSOs. You don't call CQ on a repeater, for instance. You either call a specific station or just ask if there is anyone listening. For example, I might say:

> *This is ZS1 Alpha November. Is there anyone on frequency?*

Note that I have not bothered to use phonetics for "ZS1". Since repeaters are mostly for local use, everyone will know that I have either a ZS1 or a ZR1 callsign. However, I still use phonetics for the "AN" to avoid confusion with, for example, ZS1AM.

If there is a station listening who wants to chat, he might reply:

> *ZS1AN this is ZS1 Bravo. Hi there, Andrew. Nice to hear you again! What have you been up to?*

The other station will know when you have stopped transmitting, as the repeater will shut off after a short delay. There is no need for ending signals like "Over".

Do not normally give signal reports when using a repeater, since you do not know what the strength of the other station's signal is to the repeater—you only know the strength of the repeater's output, which is of little interest to the other party. If someone does ask for a report, don't give an RST report but rather tell him or her in plain language that they are "loud and clear" or "some hiss" or "breaking up", whatever the case may be. If you say "full quieting" this means you are receiving a clear signal without any hiss on it.

From there on it is pretty much like a phone call. You should use a minimum of jargon, speak naturally, and remember to give your callsign on every transmission.

There are a number of special points regarding repeater QSOs:

1. The repeaters are a shared resource; they are not there for your private use. If you want to have a long conversation with one other person, change to a simplex frequency – that is, a frequency on which you communicate directly with the other station, not tying up a repeater. Long "group chats" where anyone can participate are accepted on repeaters and are quite common.

2. Leave a pause of 2 to 3 s after the end of the previous transmission before you start another over. This gap is to give anyone else who wants to join in the conversation the opportunity to do so.

3. A station that wants to join a conversation or net (network—a number of stations chatting) should wait for a pause between overs, and then just give their callsign, once, on the repeater. The next station to transmit should acknowledge the "breaker" and hand over to them as soon as convenient.

4. If you have an urgent message to pass on a repeater, wait for a pause and then say "break break" and your callsign. The next station to transmit should then hand over to you immediately.

5. Keep your over fairly short, with a maximum of two to three minutes. Don't give speeches or sermons over the repeater!

2.5 Emergency Communications and Social Responsibility

One of the prime reasons why amateur radio is given wide slices of spectrum world-wide, despite competition from commercial and government demand, is that it provides a unique emergency contingency communications capability. As amateurs, we must be proud of our ability to act as a backup when things go pear-shaped. Amateurs regularly make the news when natural disasters such as fires, hurricanes, earthquakes and floods destroy normal infrastructure, and the simple portable amateur station saves the day.

When dealing with emergency communications, common sense must prevail. With the wide variety of possible scenarios, there can be no specific fixed rules. A few principles apply:

- If you are in a position to help with emergency communications, make sure that the authorities know about your existence, as well as your capabilities. Advise them of the frequencies that you can access, and whether some of those frequencies are shared by other emergency services.
- Join an organisation such as Hamnet, which coordinates amateur radio emergency communications under the auspices of the SARL. Hamnet holds regular emergency communications exercises to iron out unforeseen problems with procedures and equipment. Participants also provide useful communications support at community events such as ultramarathons, triathlons and motor rallies.
- Occasionally see how you can cope when your mains power supply is turned off. See what you need to do to make your station useful as a disaster communications tool.

The word Mayday and the telegraphy signal di-di-di-dah-dah-dah-di-di-dit (popularly known as “SOS”, but actually one single symbol) may not be transmitted for any purpose except to indicate an emergency. If you are dealing with an emergency where human lives are at stake, feel free to use them. They provide listeners with a clear, unambiguous indication that there is an emergency, and they will probably be prepared to assist as required.

If you hear an emergency in progress, just listen. If you can contribute in some way, such as when you have access to telephone service when the other parties do not, by all means offer your assistance. Let the station handling the emergency decide whether you are needed or not. If you cannot help, do not transmit.

In some widespread emergencies, such as the 2015 Nepal earthquake, worldwide coordination was conducted via amateur radio. The stations advertised several frequencies in the 20 m band, and used them around the clock. Other amateurs were asked to stay off these frequencies. If you become aware of such activity, by all means stay off those frequencies if you are not directly involved and spread the word to others.

2.6 General Points

Remember that the purpose of these procedures is to facilitate clear, intelligible communication even under poor conditions. Do not use unnecessary jargon when plain language will do. For example, the Q code assists with effective communication in Morse code, where it might take too long to spell it out in full, or when talking to someone who has a limited command of English. On a repeater or when talking to a competent English speaker on HF, it is usually just as quick to use normal language, and more understandable.

Amateur radio is *not* the place to discuss contentious issues such as religion, politics or anything that anyone might regard as indecent. Apart from the obvious potential for conflict, it is also illegal under your licence conditions. With politics, it is probably best left well alone, even if you know the other person shares your views, as you will never know who else is listening.

Never use insulting, obscene or insulting language, even when your type might do so on the phone. In amateur radio there is no such thing as a private conversation and all amateurs have an interest in keeping our bands free of abusive and obscene language. If you are heard using unacceptable language even in a “private” conversation with someone who does not mind your language, other amateurs will report you to the authorities and demand that your licence be revoked.

By law you may not interfere with other QSOs in any way. If you think that someone else is hogging the repeater, by all means point this out politely to him or her. Do not be tempted to respond with jamming, as you will probably eventually lose your licence.

The bands allocated to amateurs are divided into segments for different uses according to the *band plan*. Typically each band has different segments set aside for CW, digital modes and phone. Some frequencies may be reserved for beacons, on which no other stations should transmit. Others may be reserved for particular purposes, such as satellite use or inter-continental DX. Although in most cases it is not a legal requirement to observe the band plans, courtesy to other operators should be sufficient reason to do so.

2.7 Keeping a Log

Keeping a log is a legal requirement, with certain exceptions. Most amateurs keep a log even when it is not required. An old logbook is a wonderful source of nostalgia, and you may want to use some VHF or mobile contacts for an award one day.

Time keeping is an important element of logging. For amateur radio purposes, the use of UTC is recommended. UTC is the French abbreviation for Universal Coordinated Time. It is practically the same as GMT (Greenwich Meridian Time), which was the British standard time that served as an international standard for centuries.

In South Africa, we use South African Standard Time (SAST). UTC is two hours behind SAST. When it is 14:30 in Johannesburg, it is 12:30 UTC. SAST and UTC are also referred to as B and Z time respectively. 14:30B = 12:30Z.

Be careful when operating close to midnight. 01:00B is 23:00Z on the previous day!

The advantage of UTC is that it is the same everywhere on earth. At 16:30Z, it is 16:30Z everywhere on earth. Even though it may be 11:30 local in New York, 08:30 in San Fransisco, 01:30 the next day in Tokyo and 06:30 the following day in the Line Islands, operators in all these locations understand that it is 16:30Z.

There are many software packages that can keep your log. Some of them can send Morse code and digital signals, turn your rotator, control your radio, show you world maps and print QSL labels. However, you can also use paper logs, so that you don’t need to turn on your computer all the time. Some operators keep paper logs, later entering the information into a computer to take care of the routine admin chores.

You can make your own logbook. Just record your name and address on the front page, to meet the legal requirements for a log. Then make a logsheet that contains at least the required information:

Date *2016*	Time on UTC	Callsign	RST sent	RST rcvd	Freq [kHz]	Mode	Pwr [W]	Time off UTC	Remarks
02-04	*06:14*	*VP8SGI*	*599*	*599*	*3503*	*CW*	*500*	*06:14*	*S Georgia*
	07:12	*VP8SGI*	*599*	*599*	*10115*	*RTTY*	*300*	*07:12*	*QSL via OQRS*
	09:18	*ZS4TX*	*53*	*58*	*7120*	*SSB*	*100*	*09:18*	*Bernie Bloem*
02-05	*02:15*	*PJ2T*	*339*	*559*	*1832*	*CW*	*1000*	*02:17*	*Op K8ND*
	10:00	*G3XTT*	*599*	*459*	*24895*	*CW*	*5*	*10:11*	*Don nr London*

A landscape page is probably better to allow more room to write all the necessary information. You may decide to add other columns, such as which antenna was being used, or to leave more room in some columns, such as the Remarks column.

2.8 Exchanging QSL Cards

QSL cards have a long tradition within amateur radio. In the early days, people used to exchange simple postcards to confirm contacts.

These days, most stations use pre-printed cards, onto which the information of a specific contact can be written. Many stations now use self-adhesive labels to apply the information to the card.

Cards can be exchanged via airmail. Addresses for most stations can be found in national databases, or in generic voluntary-participation databases such as QRZ.com. Many sought-after stations use other volunteers to do their QSLing for them. These selfless individuals, known as *QSL Managers*, handle all the paperwork for the common good and respond to incoming requests, often paying the expenses too. If a station has a manager, do not send the card to the station you worked. You almost certainly will not get a response.

If you want a reply from a station, it is customary to include a Self-Addressed Stamped Envelope (SASE) to cover the postage. If you do not have stamps from the country concerned, use a Self-Addressed Envelope (SAE) with return postage. Postage normally comes in the form of International Reply Coupons (IRCs). Some operators use US dollar bills, colloquially known as “green stamps”. In some countries, postage is expensive, and several IRCs or green stamps may be required.

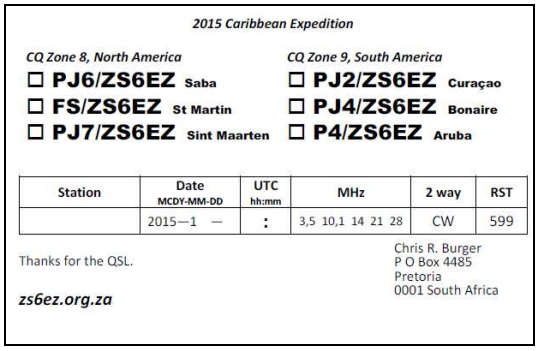
If you take the cost of two envelopes, postage, return postage and QSL printing cost into account, airmail becomes expensive. To alleviate this expense, most countries have QSL bureaux. A QSL bureau handles QSL cards in bulk. Members send their cards to the bureau, where they are merged with cards from other members and bulk-shipped to foreign bureaux. Once they arrive there, they are sorted and distributed to members, again in bulk. Bureaux take a long time—it is not uncommon for a card to arrive after a few years—but they are easy to use and economical. Apart from saving on the cost of postage, you also do not have to address the cards individually.

The SARL has such a bureau, which distributes cards to and from SARL members. They recommend that you should keep some large SASEs on file, so that your incoming cards can be regularly shipped to you. Although the use of the bureau is a benefit of SARL membership, there is a limit to the number of cards a member can send annually. Heavy users have to pay a mass-based fee above the basic quota. You should just deliver the cards to the bureau sorted by the foreign bureaux for which they are intended. If you do not understand how to sort the cards, simply deliver them in alphanumeric order (0 to 9, A to Z) by callsign.

Once you become active on the air, you should consider having some cards printed. Cards range from very basic to very fancy.

QSL cards are 90 x 140 mm, using stock of between 190 and 250 gsm. These standards facilitate handling by the bureau system, but also make the storage of cards much easier.

An example of a QSL card is shown below. This particular card can be used for several different locations. Most cards would only feature a single callsign and a single location.



2015 Caribbean Expedition

CQ Zone 8, North America

☐ **PJ6/ZS6EZ** Saba
☐ **FS/ZS6EZ** St Martin
☐ **PJ7/ZS6EZ** Sint Maarten

CQ Zone 9, South America

☐ **PJ2/ZS6EZ** Curaçao
☐ **PJ4/ZS6EZ** Bonaire
☐ **P4/ZS6EZ** Aruba

Station	Date MCDY-MM-DD	UTC hh:mm	MHz	2 way	RST
	2015—1 —	:	3,5 10,1 14 21 28	CW	599

Thanks for the QSL.

Chris R. Burger
P O Box 4485
Pretoria
0001 South Africa

zs6ez.org.za

A QSL Card (reproduced half size)

Some cards feature pictures or interesting information about the operator or the area where the station is located. A block structure like the one shown is recommended, as it makes the information about the contact easy to read and understand. A good design approach is to have a simple block design with QSO details on one side, with the picture and story on the other. This way, you can achieve a spectacular effect while still maintaining clarity and simplicity where it counts. The QSO side must contain all information, so that the person handling the QSL does not need to turn the card over.

You can design your own card. However, seek guidance from an experienced bulk QSLer to ensure that you don't forget something. There are vendors that print QSL cards at reasonable prices. There is even a service that will print your cards, prepare labels from log files you send and then ship the cards to bureaux world wide.

It is common courtesy to confirm a QSO if requested. You should upload all your logs to the Logbook of the World (LotW). LotW provides instant confirmation of contacts for awards purposes. Once you and the other station have uploaded your logs, both can use LotW to prove that the contact really took place. Participation is completely free to all participants. Only when one of the parties wants to use LotW to prove a contact for an award, does that party have to pay a minimal charge.

You should also answer all incoming QSL requests. If sufficient postage is provided, use airmail. If not, you may elect to respond via the bureau to save cost.

Finally, there is an increasing trend towards Online QSL Request Systems. Instead of sending them an envelope with your card and a request, you request the card online. ClubLog is the most popular platform for OQRS. Generally, you can request a bureau card free of charge, or a direct card with postage. ClubLog includes a Paypal-based payment mechanism. OQRS is a bargain, as it saves time, cost and risk relative to a direct request.

Revision Questions

1 To prevent annoying or jamming other users when tuning up your transmitter, initially tune:

a. On a harmonic outside the band.
b. Directly into an antenna.
c. Into a dummy load.
d. Directly into a dipole.

2 Amateur band plans are formulated and should be observed because:

a. They are mandatory.
b. They are governed by international regulations.
c. They are intended to aid operating and help to avoid congestion.
d. They are there for Novice use.

3 The term CQ is used to:

a. Call for a contact with another amateur station.
b. Terminate a conversation.
c. Interrupt a conversation.
d. Make a test transmission.

4 Before making a CQ call on any frequency one should:

a. Send a 1750 tone burst.
b. Keep giving your callsign repeatedly.
c. Listen on the frequency before and if clear, commence to call.
d. Give your callsign three times.

5 Immediately prior to transmitting, a licenced operator should always:

a. Check earthing.
b. Check antennas.
c. Check power supplies.
d. Listen to check whether the frequency is clear.

6 To ensure the calling station's callsign is clearly identified when inviting a contact, the caller should:

a. Speak slowly and clearly.
b. Speak very quickly.
c. Use maximum speech compression.
d. Use the highest frequency.

7 When calling another station, it is accepted practice to:

a. Give your callsign first and then the station being called.
b. Use only your callsign.
c. Give the callsign of the other station first, followed by your own callsign.
d. Use the callsign of the other station once only.

8 When signing off with another station at the end of a contact, it is accepted practice to:

a. Give your callsign first and then the other station's.
b. Give the other station's callsign after your callsign at the end.
c. Don't use the other stations callsign or yours but say "over and out".
d. Give the other station's callsign first and your callsign last.

9 When replying to another station, how often should the callsign of the station being called normally be given?

a. Once.
b. Twice.
c. Three times.
d. Four times.

10 Once having established contact with another station on a calling frequency, it is good practice to:

a. Continue the contact on the same frequency.
b. Move to another frequency and have a QSO.
c. Invite others to join you on the same frequency.
d. Be objectionable to all other stations calling.

11 When two stations are in QSO you should:

a. Butt into the conversation without knowing what they are discussing.
b. Listen first and after finding out the gist of the QSO ask to join and start talking about something else.
c. Butt in and start an argument about another subject.
d. Listen first and if you can contribute to the QSO, ask to join and add what you can to stimulate further discussion.

12 Before you come on the air for the first time you should:

a. Know all the procedures used on CB and use them to the full.
b. Use all CB terms even if they do not apply to Amateur Radio.
c. Learn basic amateur radio procedures and terms first and only then venture on the air.
d. Learn all commercial radio terms and use them.

13 When you call CQ for the first time and do not get a reply, you should:

a. Move up or down the band and call every few kHz.
b. Call again and again on the same frequency.
c. Change to another band.
d. Listen around the band to see if there are other stations active before calling CQ and call a few times before quitting.

14 The subject matter for any discussion on amateur radio, should include:

a. Politics, religion and sex.
b. Discuss offensive matters.
c. Use indecent language as often as possible.
d. Matters of mutual interest and of a personal or technical nature in a relaxed and dignified manner.

15 Which one of the following is correct using telephony to make a South African contact?

a. CQ CQ CQ. This is Zulu Sierra six Zulu Zulu Zulu calling, Zulu Sierra Six Zulu Zulu Zulu calling CQ and standing by.
b. CQ CQ Zulu Sierra Six Zulu Zulu Zulu standing by.
c. CQ DX CQ DX CQ DX this is Zulu Sierra Six Zulu Zulu Zulu.
d. CQ CQ This is Zulu Sierra Six Zulu Zulu Zulu. CQ Zulu Sierra Six Zulu Zulu Zulu, over.

16 Which one of the following is *not* correct?

a. It is important to speak clearly and not too fast when the other person cannot speak the same language as you.

b. Q codes should only be used on Phone when the other operator is inept in English.

c. Ham jargon and slang should be used whenever possible to confuse unlicenced listeners.

d. Avoid the use of "we" when I is meant.

17 If a station is calling "CQ Europe" you should:

a. Call him anyway.

b. If he does not answer your ZS call, curse him and accuse him of being anti South African.

c. Wait and see if he gets replies from Europe and if not, wait to hear what area he calls next, and so on until he calls CQ Africa.

d. Blow a trumpet or musical instrument to attract his attention.

18 When operating on any Amateur Radio band one should:

a. Operate wherever is convenient and unoccupied.

b. Use Lower Sideband in the Upper Sideband portion.

c. Follow the accepted Band Plan for the band being used.

d. Use CW in the phone portion if the band is clear.

19 When operating on High Frequency bands, it is good practice, after contacting a station initially, to:

a. Go straight ahead with the conversation.

b. Exchange signal reports some time during the conversation.

c. Exchange signal reports and ascertain that signals are suitable for a contact before proceeding with extended dialogue.

d. Move slightly off frequency to enable the other parties to hear better.

20 When conditions are good and signals are strong operators should:

a. Increase power to the maximum permissible by regulation.

b. Increase power and use full compressor facility.

c. Use only sufficient power to make a good contact

d. Increase power to the maximum capability of the equipment.

21 Before transmitting, which of the following procedures is not correct?

a. Check that the antenna system is in proper order.

b. Check that there is no undue reflected power on the antenna.

c. Check that the correct frequency is to be used.

d. Assume that the last settings of transmitter controls is suitable.

22 A signal report of 599 is given when a received signal has:

a. A poor signal strength with a good CW tone.

b. A good signal strength but a poor CW tone.

c. Totally unreadable CW.

d. A perfectly readable, strong and clear tone signal.

23 In the RST code, the T is for:

a. Temperature.

b. Tone.

c. Time of transmission.

d. Transmitter type.

24 A readability report of 2 would indicate:
a. Unreadable.
b. Only readable with considerable difficulty.
c. Readable with only slight difficulty.
d. Perfectly readable.

25 The S report in the RST code is obtained from:
a. The apparent strength of the received signal.
b. The speed at which CW is sent.
c. The level of interference on the band.
d. The indication on the receiver's S-meter.

26 A 59 report is commonly given to stations who:
a. Generate poorly readable signals.
b. Are unreadable.
c. Put in good strong, well-understood signals.
d. Send poor CW.

27 The term "5 and 9" used to describe a signal, is in which code?
a. Q code.
b. RST code.
c. Morse code.
d. Colour code.

28 The use of the Q code is primarily to:
a. Stop unlicenced listeners from understanding transmissions.
b. Save transmitting power.
c. Ensure effective communication.
d. Use sidebands.

29 The Q code for "stand by" is:
a. QRM.
b. QRN.
c. QRS.
d. QRX.

30 QRP is taken to refer to:
a. Close down.
b. Address is.
c. High Power.
d. Low Power.

31 QRT is taken to mean:
a. Closing down the station.
b. Standing by.
c. Fading due to propagation variations.
d. Low power.

32 "Shall I decrease power?" may be transmitted as:
a. QRP?
b. QRT?
c. QSP?
d. QTR?

33 "What is my exact frequency?" may be transmitted as:
a. QRG?
b. QRI?
c. QRU?
d. QSP?

34 The correct Q code for "change frequency to" is:
a. QSR
b. QSX
c. QSY
d. QTH

35 What is the correct Q code for "what is your location?"?
a. QRP?
b. QSP?
c. QSY?
d. QTH?

36 QRM means:
a. I am inundated with static.
b. I am being interfered with by another station.
c. I am going to do a musical transmission.
d. I need more modulation.

37 QRT is defined as:
a. I am going to send now.
b. I am going to stand by.
c. I am going to close down.
d. I am waiting for your message.

38 Which is the correct Q code for "shall I stop sending?"?
a. QRK?
b. QRL?
c. QRT?
d. QRV?

39 Which is the correct Q code for "when will you call me again?"?
a. QRH?
b. QRX?
c. QSB?
d. QSD?

40 Which is the correct Q code for "are my signals fading?"?
a. QRH?
b. QRX?
c. QSB?
d. QSD?

41 Which is the correct Q code for "are you ready?"?
a. QRG?
b. QRK?
c. QRL?
d. QRV?

42 Which is the correct Q code for "can you acknowledge receipt?"?

a. QRK?
b. QRL?
c. QRV?
d. QSL?

43 Which is the correct Q code for "shall I send more slowly?"?

a. QRK?
b. QRP?
c. QRS?
d. QRV?

44 You switch your radio set on and all you hear is a station's callsign in telephony.

a. You call "QRA?"
b. You call "QRZ?"
c. You call "What is your callsign?"
d. You listen until you understand what is going on, before you start transmitting.

45 You are a new amateur and you hear all sorts of phrases being used by other amateurs.

a. You follow suit and use these expressions.
b. You accept them as correct and acceptable.
c. You add to the vocabulary new words that you make up.
d. You use plain language with normal meanings.

46 Which is the correct phonetic spelling of the word "plug"?

a. Peter London Union Germany.
b. Papa Lima Uniform Golf.
c. Pope Lima Uniform Golf.
d. Power Lima Uniform Golf.

47 Which of the following is incorrect usage of the phonetic alphabet?

a. Bravo.
b. Sierra.
c. America.
d. India.

48 The correct way to say "P" on the radio is (with the emphasised syllable underlined):

a. Peter.
b. Peter.
c. Papa.
d. Papa.

49 The correct way to say "J" on the radio is (with the emphasised syllable underlined):

a. Japan.
b. Japan.
c. Juliett.
d. Juliett.

50 Which of the following is the correct phonetic spelling for the word "ship"?

a. Sugar Hotel Item Papa.
b. Santiago Honolulu India Papa.
c. South Hotel India Papa.
d. Sierra Hotel India Papa.

51 Callsigns using phonetics should be given:

a. At the end of every transmission.
b. On the first contact with a station.
c. At the beginning of each transmission.
d. Regularly during the contact.

52 "Coil", using the international phonetic alphabet, would be spelled as:

a. Charlie, Ocean, Italy, Lima.
b. Charlie, Oscar, India, Lima.
c. Colin, Oscar, Indonesia, London.
d. Colin, Oscar, India, London.

53 Which of the following uses the International Phonetic alphabet?

a. Boston, Uniform, Golf.
b. Bravo, Union, Gold.
c. Berlin, Uncle, Golf.
d. Bravo, Uniform, Golf.

54 Which of the following is correct, using telegraphy, to solicit an overseas contact?

a. CQ CQ CQ de ZS1XYZ ZS1XYZ ZS1XYZ.
b. CQ DX CQ DX CQ DX de ZS1XYZ ZS1XYZ ZS1XYZ K.
c. CQ DX DX DX de ZS1XYZ AR.
d. CQ DX de ZS1XYZ ZS1XYZ KN

55 When using Morse Code initially:

a. Send CQ and your callsign at a very fast speed.
b. Send your CQ and callsign at the maximum speed that you can receive.
c. Send your CQ and callsign a little slower than the maximum speed that you can comfortably receive.
d. Send AR first and then the CQ call.

56 Which mode do repeaters normally operate on?

a. AM
b. FM
c. SSB
d. CW

57 To reduce false triggers by interference, some repeaters are activated by:

a. A tone burst.
b. Any signal on the input frequency.
c. Any signal on the output frequency.
d. Remote control.

58 Continuous operation of a repeater by one station is:

a. Desirable.
b. Impossible.
c. Dangerous.
d. Inconsiderate.

59 The main purpose of a terrestrial repeater is to:

a. Increase satellite coverage.
b. Increase the range of mobile stations.
c. Increase the range of fixed stations.
d. Minimise contacts by pedestrian stations.

60 A net is taking place on 2 m. You should:

a. Call CQ on that frequency during a break in transmissions.
b. Listen for a while and then butt in even if you cannot contribute to the discussion.
c. Wait for a break in transmission, then call in and wait to be called in.
d. Whistle or use a musical instrument to attract attention.

61 When using a repeater on VHF, it is good practice to:

a. Use simplex and tell the other stations they are weak and you don't hear them at all.
b. Use maximum power and call until someone answers.
c. Use the duplex mode, and call on the input frequency and listen on the output frequency.
d. Use repeater reverse and hope for the best.

62 When using a repeater you should give:

a. A signal strength report to other stations.
b. Request a RST signal report on your signal.
c. Give RST signal reports to other stations.
d. Report that you are copying loud and clear.

63 When using a repeater it is correct procedure:

a. To pause between overs to permit another station to break in.
b. To transmit immediately after the station in the contact turned it over to you, to prevent unwanted stations interrupting your conversation.
c. To pause for a considerable time before replying to the other station in the contact.
d. To monopolise the repeater to prevent others from using it.

64 It is good practice when using a repeater:

a. To use an inefficient antenna.
b. To use an amplified microphone with a speech processor.
c. To use a radio set that overdeviates.
d. To be polite and allow other stations to join into the conversation.

65 When you wish to pass an urgent message over a repeater that is in use, you should:

a. Press the microphone switch and shout that you are in a hurry.
b. Whistle continuously to draw attention.
c. Wait until the end of the over, identify yourself, and announce that you have an urgent message.
d. Press the microphone switch and wait until both stations become silent and then take over and pass your traffic.

66 When using a repeater, you are told that your signal is breaking up and unreadable. You then:

a. Tell the other station that there is nothing wrong with your set.
b. Sign clear until you get into a better position and can access the repeater correctly.
c. Ask someone to relay your unimportant message.
d. Irritate everyone by asking for repeats of messages.

67 When using a repeater one should always:

a. Keep the overs as longs as you feel like.
b. Discuss subjects including politics, sex and religion.
c. Keep the overs short so as to allow other users access.
d. Access the repeater without giving your callsign.

68 The recommended practice is to use:

a. SAST for local contacts and UTC for DX contacts.
b. SAST for all contacts.
c. UTC for all contacts.
d. SAST in the log and UTC on QSL cards.

69 When it is 12:30B in Johannesburg, it is:

a. 10:30 local time in New York.
b. 10:30Z in New York.
c. 12:30 local time in New York.
d. 12:30Z in New York.

70 A good QSL card:

a. Has a standard size of 90 x 140 mm.
b. Is printed on standard stock of 190 to 250 gsm.
c. Has the QSO information in block format.
d. All of the above.

71 The fastest way to obtain a QSL card for a contact that you would like to confirm is:

a. SASE.
b. OQRS.
c. LotW.
d. Bureau.

72 The cheapest way to obtain a QSL card for a contact that you would like to confirm is:

a. SASE.
b. OQRS.
c. LotW.
d. Bureau.

73 The fastest way to confirm a contact is:

a. SASE.
b. OQRS.
c. LotW.
d. Bureau.

Chapter 3: Basic Electrical Concepts

3.1 Atoms and Electrons

The matter that we interact with every day consists of atoms. The term “matter” includes solid objects like desks and computers, liquids like water and gasses like the air we breathe. Atoms are tiny and invisible to the naked eye and were once thought to be the ultimate indivisible constituents of matter, but we now know that they are themselves made up of various sub-atomic particles. For the purposes of this discussion, atoms can be thought of as a very small central nucleus that is surrounded by a cloud of electrons. Electrons are not simple particles like miniature planets surrounding the nucleus, but are “smeared out” in space so that even a single electron can form a cloud around a nucleus.

The nucleus consists of one or more protons, accompanied by one or more neutrons. Protons are positively charged particles, while neutrons are uncharged (electrically neutral). The overall charge of a nucleus is always positive, from the positively charged protons. Electrons are negatively charged, and since opposite charges attract, the negatively charged electrons are attracted to the positively charged nucleus, which is what makes the electrons stay close to the nucleus.

Point to remember: *Opposite charges attract, like charges repel.*

Of course, since like charges repel, you might ask what stops the positively charged protons in the nucleus from bursting apart and destroying the atom. The answer is that another force called the “strong nuclear force” holds the nucleus together. The strong nuclear force is stronger at the very short distances characteristic of an atomic nucleus than the repulsive electromagnetic force between the positively charged protons.

Visible amounts of matter contain huge numbers of atoms. For example, a copper cube 1 mm on each side would weigh less than one hundredth of a gram, but would contain about 85 000 000 000 000 000 000 atoms!

3.2 Conductors and Insulators

In some materials, such as copper, some of the electrons are not very strongly bound to their nuclei. These electrons are free to move around in the material, as long as other electrons replace them when they move. If they were not replaced, the area they left would have more protons than electrons, giving it an overall positive charge. This charge would attract electrons back there and make it harder for other electrons to leave, since the negatively charged electrons would be attracted by the positive overall charge.

Materials in which some of the electrons can move around relatively freely conduct electricity and are known as “electrical conductors”. Materials in which all the electrons are tightly bound to their nuclei and cannot move around do not conduct electricity and are known as “electrical insulators”.

Most metals are conductors. Silver is the best conductor of all, but too expensive for most uses. Copper is a very good conductor at a more reasonable price. Aluminium is ideal for weight-sensitive applications like overhead cables. Mercury is a good conductor that is a liquid at room temperature. Solder is an alloy, often of tin and lead, with a low melting point that is used to connect electrical components together.

Good insulators include most plastics and ceramics, glass, plexiglass, rubber and dry wood. Impure water is not an insulator, so anything wet is likely to conduct electricity, especially if you did not intend it to.

3.3 Electric Current

When we speak of a material conducting electricity, we mean that electric currents can flow through that material. But what is an electric current?

> **Definition:** *An electric current is a flow of charge.*

Any time that charge is flowing – that is, moving in a relatively consistent direction – there is an electric current. Since charge is normally associated with particles of one sort or another, a flow of charge usually entails a flow of charged particles, such as electrons. The particles that carry the charge are known as "charge carriers".

The size of an electric current is expressed in ampere, named after the French physicist André-Marie Ampère (1775-1836) who was a pioneer in the study of electricity. The official abbreviation is "A", but the slang term "amp" is widely used.

When an electric current flows through ordinary conductors like copper, the charge carriers are electrons, so the flow of electric current corresponds to a flow of electrons. However, because electrons are negatively charged, electrons flowing from left to right through a wire would constitute a *negative* current flowing from left to right in the wire. This current could also be described as a *positive* current flowing in the opposite direction, from right to left in this case. Electric current is generally considered to flow in the *opposite* direction from the electrons that carry it!

Whenever someone refers to just an "electric current" you should assume that they are talking about a conventional current, so if the charge carriers are negatively charged particles like electrons, the direction in which the current flows will be the opposite direction to the flow of charge carriers.

You can imagine a (conventional) electric current flowing in a wire to be similar to water flowing through a pipe. The magnitude (size) of the current would correspond to the rate of water flow through the pipe. A suitable unit for water flow might be ℓ/s (litres per second).

3.4 Electric Potential

Having established that an electric current is a flow of charge, the next question is: What makes the charge flow? The answer is electric potential difference. Since unlike charges attract, if you apply a positive potential to one end of a conductor and a negative potential to the other end, loosely bound electrons in the conductor will be attracted towards the positive potential and repelled by the negative potential, causing electrons to move from the negative end to the positive end. In other words, a conventional current will flow from the positive end of the conductor to the negative end. As the name indicates, electric potential difference is always measured between two points.

> **Definition:** *The electric potential difference between two points is the amount of energy that it would take to move one unit of charge from the point of lower potential to the point of higher potential.*

Since energy is measured in joule and charge in coulomb (abbreviated J and C respectively), the unit of electric potential is joule per coulomb (J/C). This unit is named the "volt" with the abbreviation "V", named after the Italian scientist Count Alessandro Volta (1745-1827) who invented the battery. Electric potential difference is commonly referred to simply as "voltage".

Electric potential difference is analogous to the pressure that a pump creates in the water it pumps through a pipe. The higher the pressure (voltage), the greater the quantity of water flowing through the pipe per second (current).

3.5 Units and Abbreviations

If you measure or calculate the amount of something, you usually need to specify the unit of measurement. For example, saying that something weighs "10" does not mean much unless you specify the unit of measurement—10 grams, or 10 kilograms, or 10 milligrams.

The units of measure used in this course are the standard SI units that are used universally, except in a few uncivilised countries. Each unit has a name, like "volt" or "ampere", and a corresponding abbreviation, like "V" for volt and "A" for ampere. This notation saves time when writing quantities—for example a current of "10 A" rather than "10 ampere".

There are also a number of standard prefixes, which are used to indicate quantities a thousand or a million or more times bigger or smaller than the basic unit. For example, the prefix "milli" which is abbreviated "m" means "one thousandth of", so one milligram—written as "1 mg"—means one thousandth of a gram. The following prefixes are widely used in electronics:

Prefix	Abbreviation	Scale Factor	Scientific Notation
pico	p	÷ 1 000 000 000 000	10^{-12}
nano	n	÷ 1 000 000 000	10^{-9}
micro	μ	÷ 1 000 000	10^{-6}
milli	m	÷ 1 000	10^{-3}
kilo	k	x 1 000	10^{3}
mega	M	x 1 000 000	10^{6}
giga	G	x 1 000 000 000	10^{9}

Note that the case in which a prefix abbreviation is written (capital or lower case) is important. Mm and mm are not the same thing.

Names of units are generally written in lower case: coulomb, hertz or ampere. The abbreviation for the unit is associated with a specific case: C, Hz, A. s is a second, the unit for time. S is siemens, the unit for conductance.

Finally, there are some multiples that can be used, but are not preferred:

Prefix	Abbreviation	Scale Factor	Scientific Notation
centi	c	÷ 100	10^{-2}
deci	d	÷ 10	10^{-1}
deka	D	x 10	10^{1}
hecto	h	x 100	10^{2}

3.6 Scientific Notation

The column headed "scientific notation" may not be familiar to you. Because scientists work with very small and very large numbers, it would be inconvenient for them to have to keep writing many zeroes after the large numbers, or a decimal comma and many zeros before the small numbers. So they use the fact that multiplying by ten to the power of any positive number effectively adds that many zeros at the end of the number. So for example the speed of light is about 3×10^{8} m/s which means "3 followed by 8 zeros", or 300 000 000 m/s. This quantity could also be expressed as 300 000 km/s or 300 Mm/s.

Another way of thinking of this is that it is equivalent to moving the decimal point 8 places to the right, and introducing as many zeros as are necessary to do so. This trick is helpful when the number already has a decimal comma, for example "2,998 x 10^8". You can't simply think of adding zeros, since adding eight zeros to "2,998" would give you "2,99800000000" which represents the same number, only to a greater precision. However, if you instead think about moving the decimal comma eight places to the right and adding zeros as necessary you get the correct result, which is 299 800 000. The power of ten—in this case, 8—is known as the "exponent". Most scientific calculators have a key marked "E" or "Exp" which is used to enter numbers in this format.

Similarly, a negative exponent means you move the decimal point that number of places to the *left*, again filling in zeros as required. So for example, 1,6 x 10^{-19} is equivalent to 0,000 000 000 000 000 000 16, a very small number indeed. If you were wondering, it is the charge on a single electron, in coulomb.

3.7 Number Formats

In this document, we follow the ISO convention when writing numbers. Decimal dividers are written as a comma. Thousands are separated by a space. Examples:

3,1415 is read as "three comma one four one five", and is a just over 3.
12 345,678 901 is twelve thousand three hundred and forty five comma seven.
1,5 x 10^{-5} is 0,000 015 (zero comma zero zero zero, zero one five)

Numbers with up to four digits are simply written as a unit: 7 or 27 or 276 or 2762 or even 3,1415. Numbers with more than four digits are grouped in groups of three digits: 12 475 or 13 742 186 or 3,141 592 653 589 793 238 462 643 383.

Really large and really small numbers will be written using SI prefixes or powers of 10 (e.g. 10 pF = 10 x 10^{-12} F = 1 x 10^{-11} F).

Summary

This module has introduced the concepts of electric charge, electric current, and electric potential. You have seen how the atomic structure of materials allows electric currents to flow through some materials, which we call conductors, but not through others, which we call insulators.

Electricity is obtained by converting other forms of energy into electricity. All energy being used on earth ultimately comes from the sun.

You have learnt the meaning of the prefixes that are used to scale units by powers of ten, and to understand numbers written in scientific notation.

Revision Questions

1 One of the following is an electrical insulator:

a. Silver.
b. Aluminium.
c. Copper.
d. Mica.

2 One of the following is an electrical conductor:

a. Mica.
b. Ceramic.
c. Plastic.
d. Copper.

3 The unit of electrical potential is the:

a. ampere.
b. amp.
c. voltaire.
d. volt.

4 A current of 15 μA is equivalent to:

a. $1,5 \times 10^{-5}$ A
b. 15×10^{-5} A
c. $1,5 \times 10^{6}$ A
d. 15×10^{6} A

5 A voltage of 20 000 V could be expressed as:

a. 20 μV
b. 20 mV
c. 20 kV
d. 20 MV

6 The charge carriers in solid copper that allow it to conduct electricity are:

a. positively charged copper ions.
b. negatively charged copper ions.
c. positively charged electrons.
d. negatively charged electrons.

7 Conventional current flows:

a. in the same direction as electrons are moving.
b. in the opposite direction to the flow of electrons.
c. at right angles to the flow of electrons.
d. from negative to positive.

8 An electric current always consists of a flow of:

a. electrons.
b. neutrons.
c. protons.
d. charge.

9 Electricity is obtained from:

a. Splitting atoms into protons and electrons.
b. Converting other forms of energy into electricity.
c. Electricity plants.
d. Electric blankets.

Chapter 4: Resistance and Ohm's Law

4.1 Resistance

In the last module we learnt that an electric current is a flow of charge that is caused by a potential difference between two points. We also saw that the greater the electric potential between two points joined by a conductor, the greater the current that would flow through the conductor.

However, the electric potential difference between two points is not the only factor that determines the size of the current flowing between them. The current flow is also affected by a quality of the conductor, known as its resistance. The resistance of a conductor can be thought of as being the extent to which it resists the flow of current. The greater the resistance of a conductor, the lower the current that will flow through it for a given potential difference. Conversely the lower the resistance of the conductor, the greater the current that will flow through it for a given potential difference.

The German physicist Georg Ohm (1789-1854) discovered that the potential difference across a conductor is proportional the current that flows through the conductor. In other words, if the current flowing through a conductor doubles then the voltage across that conductor will double. Conversely if the current through the conductor halves then the voltage across the conductor will also be halved.

Mathematically, this relationship can be expressed by saying that the voltage across the conductor is equal to the current through the conductor multiplied by some constant (for that particular conductor). Ohm called this quantity the "resistance" of the conductor,

voltage = current x resistance

This relationship is known as "Ohm's Law".

The unit of resistance is the ohm. The abbreviation for the ohm is the Greek capital letter omega, which is written as Ω. A conductor has a resistance of one ohm if the application of a potential difference of one volt across the conductor causes a current of one ampere to flow through the conductor.

Resistance may be thought of as the opposition to the flow of electric current through a conductor or electric circuit. Returning to our analogy with water flow, a thin pipe or a fine filter will exhibit a lot of resistance, while a thick pipe will exhibit low resistance.

4.2 Symbols in Mathematical Equations

In order to save time when writing out equations, it is common practice to use symbols to represent quantities rather that writing out the full names of quantities like "voltage" and "resistance" every time.

Certain symbols are commonly used to represent particular quantities. For example, "V" is commonly used to represent an electric potential difference (voltage), and "R" is usually used to represent a resistance. A current is usually represented by "I". Unfortunately, these symbols do not always correspond to their English names.

Using these symbols instead of the full names of the quantities, Ohm's Law us usually written as:

Ohm's Law: $V = I R$

Note that the multiplication sign between "I" and "R" is also omitted. In mathematics, when two symbols are written next to each other it is assumed that they are to be multiplied together.

This form of Ohm's Law is convenient if you know the current flowing through a conductor and the resistance of the conductor, and want to calculate the electrical potential (voltage) across the conductor. It shows that you can calculate the voltage by multiplying the current by the resistance.

For example, if a current of 5 A is flowing through a conductor with a resistance of 2 Ω, the voltage across the conductor can be calculated by replacing the "I" with 5 A and the "R" with 2 Ω in the equation for Ohm's Law, giving

$$V = 5\ A \times 2\ \Omega$$
$$= 10\ V$$

Note the somewhat confusing use of "V" both as the symbol for voltage and also as the abbreviation for the unit "volt". In this equation, the V on the left hand side (before the equals sign) is the symbol for electric potential. The V after the number 10 is the abbreviation for the unit, volt. The two meanings are not the same and you should take care not to confuse them. You should be able to work out the correct meaning from the context in which the "V" appears.

The symbol E is also used for electric potential. So you may see Ohm's Law written as $E = I R$ instead of $V = I R$.

4.3 Rearranging Ohm's Law

We now know how to calculate the voltage when we have the current and the resistance. However, Ohm's Law can also be used to find either the current or the resistance if both the other quantities are known. This is done by using simple algebra to rearrange Ohm's Law as follows:

$$V = I R$$

By dividing both sides by I you get

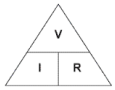
$$V/I = R$$
or $$R = V/I$$

This formula can be used to calculate the resistance of a conductor given the voltage across the conductor and the current flowing through it. Similarly, if you divide both sides of the original equation by R you get

$$V/R = I$$
or $$I = V/R$$

In this form, Ohm's Law can be used to calculate the current flowing through a conductor given the electrical potential (voltage) across the conductor and the current flowing through it. You need to be able to use any of these three forms of Ohm's Law in the examination.

Hint: If you are a bit rusty with rearranging equations, you may consider using the triangle trick. This triangle shows the relationship between V, I and R. If you look at any of the three variables, you can see its relationship to the other two at a glance. See if you can obtain all the different forms of Ohm's Law shown above by using this triangle.



Summary

Ohm's Law states that the electric potential across a conductor is proportional to the current flowing through the conductor. It can be written as $V = I R$, where R is a constant of proportionality that is known the *resistance* of the conductor. Resistance may be thought of as the opposition to the flow of electric current through a conductor or electric circuit. Resistance is measured in *ohm*, with the abbreviation Ω. Ohm's Law can be used to find the electric potential across a conductor, or current flowing through the conductor, or the resistance of the conductor provided that the other two quantities are known.

Revision Questions

1 The opposition to the flow of current in a circuit is called:

a. Resistance.
b. Inductance.
c. Emission.
d. Capacitance.

2 The current through a 100 Ω resistor is to 120 mA. What is the potential difference across the resistor?

a. 120 V
b. 8,33 V
c. 83,33 V
d. 12 V.

3 The resistance value of 1200 Ω can be expressed as:

a. 12 kΩ
b. 1,2 kΩ
c. 1,2 MΩ
d. 0,12 MΩ

4 How can the current be calculated when the voltage and resistance in a DC circuit are known?

a. I = E/R
b. P = I E
c. I = R E
d. I = E R

5 A 12 V battery supplies a current of 250 mA to a load. What is the input resistance of this load?

a. 0,02 Ω
b. 3 Ω
c. 48 Ω
d. 480 Ω

6 If 120 V is measured across a 470 Ω resistor, approximately how much current is flowing through this resistor?

a. 56,40 A
b. 5,64 A
c. 3,92 A
d. 0,25 A

7 How can the oltage across a resistor be calculated when the resistance of and current flowing through the resistor are known?

a. V = I / R
b. V =R / I
c. V = I R
d. $V = I^2 R$

8 The law that relates the current flowing through a conductor to the electric potential applied across the conductor is known as:

a. Kirchoff's Current Law
b. Kirchoff's Voltage Law
c. Kirchoff's Current and Voltage Law
d. Ohm's Law

Chapter 5: The Resistor and Potentiometer

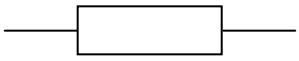
5.1 The Resistor

Electronic circuits are usually constructed from components that can be purchased at electronics outlets. One such component is the *resistor*, which is simply a conductor that has a known resistance. Resistors are available in values ranging from a fraction of an ohm to several hundred mega-ohms.

Resistors also come in different tolerances. The tolerance shows how close the actual value of the resistor is guaranteed to be to its nominal value. For example, the actual resistance of a 1 kΩ resistor with a tolerance of 5% could range from 950 Ω (1 kΩ - 5%) to 1050 Ω (1 kΩ + 5%).

Resistors also come in various power ratings. As you will see later, the power dissipated by a resistance depends on the current flowing through the resistance and the voltage across the resistance. In order to cater for different requirements, resistors are usually available in power ratings from one eighth of a watt (125 mW) to 5 W or more.

All electric components have symbols that can be used to draw diagrams showing how the components should be connected to create a particular circuit. These diagrams are known as "schematic diagrams" and the symbol for a resistor in a schematic diagram is:



Symbol for a resistor

In schematic diagrams, a plain line is used to represent a connection between two or more components, so the lines coming out of the left and right of the resistor represent its connections to the rest of the circuit. The resistor itself is the rectangle between these lines. This symbol represents a simple fixed resistance. It has two connections (represented by the lines at the top and bottom) and there is a known resistance between these connections.

In older schematic diagrams you may also see a resistor represented as a zigzag line, but we will not use that symbol.

5.2 Different Types of Resistor

Resistors come in several different types, suited to specific applications:

- Carbon Film resistors are the most common, inexpensive, general-purpose resistors. They typically have a tolerance of ±5% and power ratings from 125 mW to 2 W.

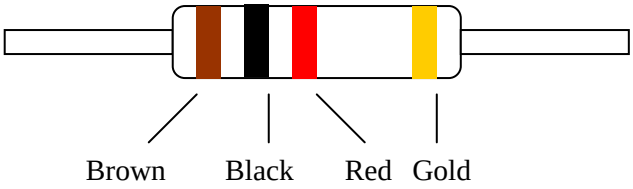
- Metal film resistors are often used when tighter tolerance is required (i.e. the resistor must be guaranteed to be closer to the nominal value). Metal film resistors typically have tolerances of ±1% or better and power ratings from 125 mW to 500 mW.

- Wire-wound resistors are used in DC applications when high power ratings are required. They are available in tolerances of ±5% or ±10% with power ratings from 2,5 W to 20 W or more. *Note that wire wound resistors should never be used in radio-frequency applications because they have unacceptably high inductance. You will learn about the effects of inductance later.*

- Resistor networks consisting of a number of resistors in various circuit configurations are supplied in packages that look like integrated circuits. They are intended for low-power applications and are especially useful when you need many resistors of the same value.

5.3 The Resistor Colour Code

Resistors are very small components, often only a few millimeters long, so if the value of a resistor (its nominal resistance, in ohms) were printed on the resistor it would be very difficult to read. So instead of printing the value onto resistors, a standard colour code is used where the value of the resistance is represented by three coloured bands, and the tolerance of the resistor by a fourth band. The following diagram represents not the circuit symbol for a resistor, but rather the physical resistor itself, showing the location of the colour bands.



From left to right the first two bands represent the first two digits in the value of the resistor. In this case, brown represents “1” and black represents “0” so the first two digits of the value are “10”. The third colour band – red in this case – represents the number of zeros that should be added after the first two digits in the value (in other words, the exponent in scientific notation). Since red represents the value “2”, two zeros must be appended to the first two digits, giving a value of 1000 Ω or 1 kΩ.

The last band, the gold one at the far right hand side, gives the tolerance of the resistor. Since gold means ±5%, the actual value of the resistor may range from 5% below the nominal value of 1 kΩ to 5% above the nominal value.

Colour	**Digit**	**Digit**	**Tolerance**
Black	0	x 1	
Brown	1	x 10	1%
Red	2	x 100	2%
Orange	3	x 1000	
Yellow	4	x 10 000	
Green	5	x 100 000	
Blue	6	x 1 000 000	
Violet	7	x 10 000 000	
Grey	8	x 100 000 000	
White	9	x 1 000 000 000	
Gold			5%
Silver			10%

For each colour the table shows you the digit that it represents when it occurs in the first two bands, the multiplier it represents when it appears in the third band, and the tolerance that it represents when it occurs in the last band.

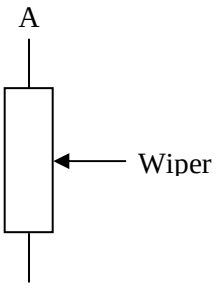
Resistors with tight tolerances, such as 1% or 2% resistors, may have an extra band in the colour code. In this case, the first *three* bands represent the first three digits of the value so that the value of the resistor can be represented more precisely. The remaining bands represent the multiplier and tolerance as before.

5.4 Expressing Resistor Values

Because resistors are very common components, a couple of shortcuts may be taken when writing resistor values. The first is that the "ohm" or Ω abbreviation for the unit may be omitted, so a 10 kΩ resistor may be referred to just as "10k". The second is that the k or M (for kilo and mega respectively) may be written where the decimal point would normally be, and the decimal point omitted altogether. So a 3,3 kΩ resistor might be written as "3k3", and a 1,5 MΩ resistor as "1M5". The character "R" is also sometimes used in place of the decimal point when there is no scale factor. A 1,5 Ω resistor might be written as "1R5".

5.5 The Potentiometer

A related component is the potentiometer, which has a variable resistance. A potentiometer is typically constructed as a circular carbon or wirewound track with a known resistance and a wiper that can be moved over the track by turning a control knob. The resistance from one side of the track to the other remains constant, but the resistance between either end and the wiper depends on the position of the control knob. The symbol for a potentiometer is shown below.



Potentiometer

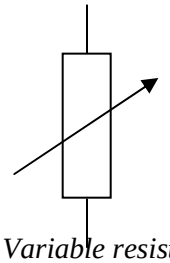
The two ends of the circular track are represented by the connections at the top and bottom of the diagram. The resistance between these points is fixed. The arrowhead represents the wiper. The three terminals are labelled "A", "B" and "W" (for "wiper") for reference in the following explanation only. These labels are not customary in most circuits.

Let us assume that the potentiometer has a value of 10 kΩ (10 000 Ω). The resistance between A and B is always 10 kΩ. When the wiper is in a central position, as represented in the diagram, the resistance between A and W would be about half of the total, around 5 kΩ, and the resistance between B and W would be the other half of the resistance, also around 5 kΩ.

Suppose we turn the control knob so the wiper is closer to A than to B. Then the resistance between A and W would be less than half, say about 2 kΩ. The resistance between B and W would be the remainder of the 10 kΩ total resistance, in this case about 8 kΩ. Similarly, if we set the wiper all of the way over to B, the resistance between B and W would be close to 0, while the resistance between A and W would be the entire 10 kΩ.

So the resistance between A and W and the resistance between B and W when added together always equal the resistance from A to B, which is the value of the potentiometer.

Potentiometers are often used as controls on electronic equipment, for example the volume control on an audio amplifier or radio receiver. There is also another symbol for a potentiometer:



Variable resistor

In this symbol, only the top and bottom lines represent connection points. The line with the arrow point does not represent a separate connection, but rather means that the resistance is variable. This typically represents exactly the same component as the more usual three-terminal symbol shown above. However, only two of the terminals are used: one side of the carbon track and the wiper. The other side of the carbon track is left unconnected, or connected to the wiper.

Although I have drawn the symbols for the potentiometer vertically, while I drew the symbol for the resistor horizontally, this was purely for convenience. Any of the symbols, like most electronics symbols, can be drawn either horizontally or vertically.

Finally, you may encounter the slang term “pot”. With the background you’ve just learned, you’ll know what it means!

Summary

The resistor is an electronic component with a defined resistance, tolerance and power rating. The tolerance is the percentage by which the actual resistance may deviate from the nominal value of the resistor. The value and tolerance of resistors is represented using the resistor colour code. The potentiometer is a variable resistor.

Revision Questions

1. A potentiometer is a:

a. Meter.
b. Variable resistor.
c. Battery.
d. Capacitor.

2. How can you determine a carbon resistor's electrical tolerance rating?

a. By using a wavemeter.
b. By using the resistor's colour code.
c. By using Thevenin's theorem for resistors.
d. By using the Baudot code.

3. Which of the resistors below (each identified by its colour coding) would be nearest in value to a 4k7 resistor?

a. Orange violet orange.
b. Yellow green red.
c. Orange violet red.
d. Yellow green orange.

4. What would the colour code be for an 820 Ω resistor, excluding the tolerance band?

a. grey red black
b. grey red brown
c. red grey black
d. red grey brown

5. What would the value of a resistor with the colour code orange orange orange be?

a. 333 Ω
b. 3,3 kΩ
c. 33 kΩ
d. 330 kΩ

6. A 10 K resistor has a gold tolerance band. The actual resistance may be:

a. From 9 000 to 11 000 Ω
b. From 9 500 to 10 500 Ω
c. From 9 800 to 10 200 Ω
d. From 9 900 to 10 100 Ω

7. A 2,2 Ω resistor might be labeled on a schematic diagram as

a. 2k2
b. 2M2
c. 2R2
d. 22R

8. The label "4M7" on a circuit diagram could refer to:

a. A resistance of 4,7 MΩ
b. A current of 4,7 MA
c. A voltage of 4,7 MV
d. Any of the above.

9. The circuit diagram for a resistor is:

a. A straight line.
b. A circle containing a zig-zag line.
c. A rectangle.
d. A triangle.

10. Which of the following types of resistor would not be suitable for radio-frequency applications?

a. A carbon film resistor.
b. A metal film resistor.
c. A wire-wound resistor.
d. A resistor network.

Chapter 6: Direct Current Circuits

6.1 Direct Current and Voltage

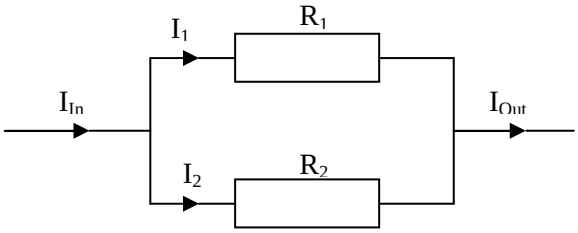
Direct current (abbreviated "DC") means a current that is flowing constantly in one direction. It is contrasted to alternating current (AC) like the mains supply, where the direction in which the current flows changes periodically, usually many times every second. Despite the apparent contradiction in terms, it is common practice to speak of a "DC voltage" to mean a constant voltage, and an "AC voltage" to mean a voltage that is reversing polarity (i.e. exchanging positive and negative terminals) periodically. Although for the moment we shall only consider DC circuits, when we come to AC circuits we will see that the same principles apply.

Remember that "voltage" is a commonly used term meaning electric potential difference. As is common practice in industry, we will use the term "voltage" rather than "electric potential difference" for the remainder of these notes.

6.2 Kirchoff's Laws

Gustav Kirchoff (1824 - 1887) formalised two very simple laws that allow us to analyze electric circuits. The first is known as Kirchoff's current law.

Kirchoff's Current Law: *At any point in a circuit where two or more conductors are joined, the sum of the currents flowing into the point is equal to the sum of the currents flowing away from the point.*



Parallel resistors

The diagram above shows two resistors connected in parallel. The arrows on the lines represent currents. A current I_{IN} flows into the circuit from the left, divides into two currents I_1 and I_2, which flow through resistors R_1 and R_2 respectively. After flowing through the resistors, the currents join again together to give I_{OUT}. For obvious reasons, this parallel arrangement is known as a *current divider*.

Note that this is not a complete circuit, as we have not shown the source of electric potential that is causing the current to flow. We must assume that there is some voltage source with its positive terminal connected to the wire on the left hand side of the diagram and its negative terminal is connected to the wire on the right hand side of the diagram in order to make the current flow.

Applying Kirchoff's current law to the point where I_{IN} splits into I_1 and I_2, we see that the sum of the currents flowing into the point – in this case there is only one current, I_{IN} – must equal the sum of the currents flowing out of the point – in this case, $I_1 + I_2$. One way to look at this is that current is a flow of charge, and charge cannot accumulate at a point, so charge must flow out of the point just as fast as it flows in.

In our analogy with a water pipe, if you put a "T" connector on a pipe then the rate at which the water flows out of the two output pipes combined must equal the rate at which the water

is flowing into the input pipe, since the water that is coming in has to go somewhere and it cannot accumulate in the T connector.

So in the diagram above we have

$$I_{In} = I_1 + I_2$$

Referring now to the point where I_1 and I_2 join together to form I_{OUT}, we can again apply Kirchoff's current law which says that the sum of the currents flowing into the point – that is, $I_1 + I_2$ – must equal the sum of the currents flowing out of the point, in this case just I_{OUT}. So this application of Kirchoff's Current Law gives us

$$I_1 + I_2 = I_{OUT}$$

Because both equations have "$I_1 + I_2$" on one side of the equals sign, we can combine them to get

$$I_{In} = I_{Out}$$

which makes sense because the charge that is flowing in on the left hand side has to go somewhere, and the only place for it to go is out the right hand side of the diagram.

The second of his laws is Kirchoff's Voltage Law. It can be formulated in two different but equivalent ways. The first formulation, which I find the most useful, is as follows.

> **Kirchoff's Voltage Law (1):** *The voltage between any two points in a circuit is equal to the sum of the voltage drops along any path connecting those points.*

This requires some explanation. Consider the circuit below:

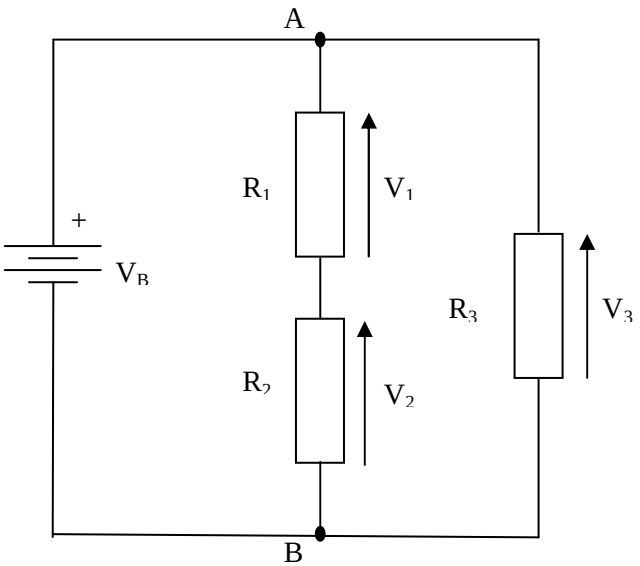


Illustration of Kirchoff's laws

The symbol on the left hand side of the diagram represents a battery. The long line always represents the positive terminal, but I have labelled it with a "+" sign to make it clear. I have also labelled the battery voltage as V_B. The battery is generating a voltage across R_1 and R_2, which are connected "in series", and across R_3, which is connected "in parallel" with R_1 and R_2.

The voltage applied by the battery will cause a current to flow through R_1 and R_2 and another (possibly different) current to flow through R_3. However, we know from Ohm's Law that when a current flows through a resistance there will be a voltage across the resistance. The voltage across a resistance is often referred to as a "voltage drop", and I have labelled the voltage drops across R_1, R_2 and R_3 as V_1, V_2 and V_3 respectively. The lines with arrowheads are used to indicate what points the voltage drop is across. Note that by convention the arrowhead points towards the positive side, which means that the arrows point in the opposite direction from the direction in which current is flowing in the circuit. In this circuit, the currents in the resistors are all flowing from top to bottom.

> **Voltage Drop:** *the potential difference across a component like a resistor caused by the current flowing through the component.*

So what does Kirchoff's Voltage Law tell us about the circuit? Consider points A and B in the diagram. Kirchoff's voltage law tells us that the voltage between points A and B is equal to the sum of the voltage drops along any path connecting A and B. If we call the voltage between A and B "V_{AB}", applying Kirchoff's Voltage law to the three different paths between A and B gives us:

$V_{AB} = V_B$ (from the path through the battery)
$V_{AB} = V_1 + V_2$ (from the path through R_1 and R_2)
$V_{AB} = V_3$ (from the path through R_3)

In other words, the *same* voltage is found across the battery, across the series combination of R_1 and R_2 and across R_3. Thinking of it in another way, the battery voltage V_B has been applied across both the series combination of R_1 and R_2 and across R_3. The concept is very simple and straightforward, and you should be able to apply it intuitively and hardly ever have to think about its formal statement as Kirchoff's Voltage Law.

A series connection, such as the combination of R_1 and R_2, is known as a voltage divider.

The second formulation of Kirchoff's voltage law is:

> **Kirchoff's Voltage Law (2):** *The sum of the voltage drops around any closed circuit is zero.*

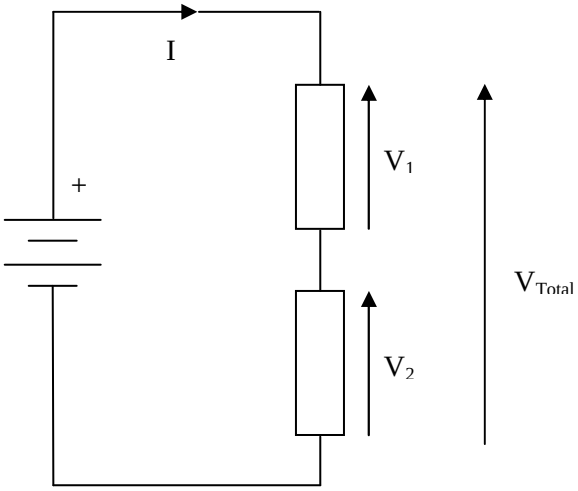
This formulation is somewhat less intuitive than the original formulation. Suppose we take a clockwise trip around the outside circuit in the diagram above, starting and ending at point A. We first go "through" the resistor R_3, and so V_3 is our first voltage drop. Staying on the outside circuit (and so ignoring R_1 and R_2), we next come to the battery. However, the voltage across the battery, V_B, is not actually a voltage *drop* because we are moving from the negative terminal to the positive terminal so the voltage is *increasing*. However, we can't just ignore it, so we instead count the battery voltage V_B as a *negative* voltage drop and add $-V_B$ to our "sum of voltage drops". Since adding the negative of a number is the same as subtracting that number we get:

sum of voltage drops = $V_3 - V_B$

However, we have already seen that V_3 and V_B are equal, so the sum equals zero and Kirchoff is happy!

6.2 Resistors in Series

Having mastered Ohm's and Kirchoff's Laws, we can use these to derive some simple and well-known results. The first is the formula for calculating the effective resistance of two resistors in series. Consider the following circuit:



Circuit with series resistors

It shows two resistors connected "in series" so that the same current flows through both of the resistors, although the voltages across each resistor may be different. The current flowing in the circuit is I, while the voltages across R_1 and R_2 are V_1 and V_2 respectively. The voltage across both resistors combined as V_{Total}. The battery is only shown for completeness, to show how the current is being made to flow in the circuit.

Suppose we want to replace the two separate resistors R_1 and R_2 by a single resistor, which will have the same effect. What value of resistor should we choose?

Note that the derivation below is provided for interest only and will not be examined. You only need to know the result that appears in italics at the bottom of this section.

From Ohm's Law,

$$V_1 = I R_1$$
and $$V_2 = I R_2$$

From Kirchoff's Voltage Law

$$V_{Total} = V_1 + V_2$$

Replacing V_1 and V_2 in this formula with the values from Ohm's Law,

$$V_{Total} = I R_1 + I R_2$$
$$= I (R_1 + R_2)$$

But this is just Ohm's Law for a resistor with the value $R_1 + R_2$. In other words, the resistors R_1 and R_2 together behave just as though they were a single resistor with the value $R_1 + R_2$. This relationship gives us the result we are looking for:

When two or more resistors are connected in series, the combined resistance is the sum of the individual resistances.

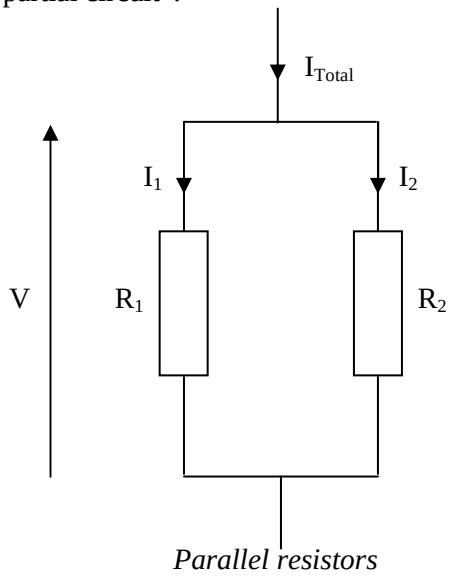
Using similar reasoning, it is easy to generalise the result to any number of resistors. Try to prove this result. You don't need Kirchoff's and Ohm's Laws, you can just use the result for two resistors and the properties of addition.

For example, if three resistors with the values 1 kΩ, 2 kΩ and 4 kΩ were connected in series the combined resistance would be 7 kΩ.

6.3 Resistors in Parallel

Another way of connecting components is to connect them in *parallel*, so the same voltage appears across each of the components although the currents through them may (and probably will) differ. Another term for parallel is "*shunt*", which is often used in high-power applications.

Consider the following circuit, which shows two resistors connected in parallel. This time the source of the potential difference has been omitted—perhaps it could be described as a "partial circuit".



The same voltage, V appears across both resistors. The currents through them are I_1 and I_2, while the total current through both resistors combined is I_{TOTAL}.

Once again the derivation is provided for interest only and is not required for the examination.

Using Ohm's Law,

$$I_1 = V / R_1$$

and

$$I_2 = V / R_2$$

According the Kirchoff's Current Law,

$$I_{Total} = I_1 + I_2$$

Substituting the values of I_1 and I_2 obtained using Ohm's Law,

$$I_{Total} = V / R_1 + V / R_2$$

Applying Ohm's Law to the whole circuit,

$$V / R_{Parallel} = I_{TOTAL}$$
$$= V / R_1 + V / R_2$$

Where $R_{Parallel}$ is the equivalent resistance of the two resistors in parallel. Dividing by V,

$$1 / R_{Parallel} = 1 / R_1 + 1 / R_2$$

This is the result we were looking for, as it shows the relationship between the value of the combined parallel resistance and the individual resistances. It is not as easy to put into words as it was for resistors in series, but I'll give it a go:

> *When two or more resistors are connected in parallel, the reciprocal of the equivalent parallel resistance is the sum of the reciprocals of the individual resistances.*

Note: The *reciprocal* of a number is *one divided by* that number. Example: The reciprocal of 2 is ½.

Of course, this leaves us with the *reciprocal* of the value we are looking for. Fortunately it is simple to convert the reciprocal of a number back into the number itself. Just calculate the reciprocal of the reciprocal and this will be the original number! For example, suppose a 220Ω resistor is connected in parallel with a 330Ω resistor. We can find the equivalent combined resistance of the two resistors in parallel as follows:

$$\begin{aligned} 1 / R_{Parallel} &= 1 / R_1 + 1 / R_2 \\ &= 1/(220\ \Omega) + 1/(330\ \Omega) \\ &= 0{,}004\ 55\ \Omega^{-1} + 0{,}003\ 03\ \Omega^{-1} \\ &= 0{,}007\ 58\ \Omega^{-1} \end{aligned}$$

So $R_{Parallel} = 1 / (0{,}007\ 58\ \Omega^{-1})$ (the reciprocal of the reciprocal!)
$= 132\ \Omega$

There is a shortcut that can be applied when all the resistances in parallel have the same value. In this special case, if the resistors all have the value *R* and there are *N* resistors connected in parallel, the equivalent resistance is *R/N*. I leave the proof of this as an exercise for the interested reader.

Returning to our water flow analogy: If you connect several pipes in parallel, the water will flow more easily and the "resistance" will decrease.

Practical Example

A "dummy load" is a high-powered resistor that can be connected to the antenna port of a transmitter. It enables the transmitter to be tested or aligned without actually having to transmit a signal. Transmitting a signal during testing when not absolutely necessary would cause interference and would be considered extremely bad manners by amateurs.

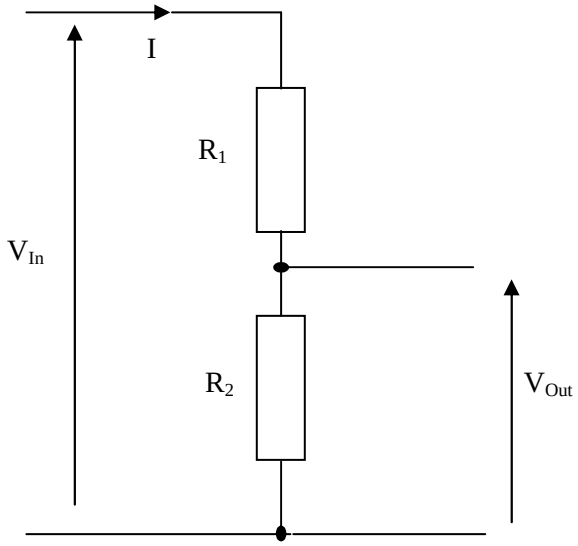
Commercial dummy loads are available but they are quite expensive. An alternative for the amateur is to make your own. Unfortunately the most commonly available suitable resistors only have a power rating of 2 W, while most transceivers will put out 100 W and would incinerate a 2 W resistor. One solution is to use fifty 2 W resistors all connected in parallel, so that each handles one fiftieth of the transceiver's power. If the resistors are each 2500 Ω (2k5) then the effective resistance of 50 resistors in parallel is 2500/50 Ω = 50 Ω, which is the correct value to match most amateur transceivers.

Remember that you will also want to shield the dummy load to prevent it from inadvertently becoming a fully functional transmitting antenna. When I built one, I did this by enclosing it in a metal baking powder tin which I chose because it has a screw-on lid. I drilled a hole in the bottom of the tin to accommodate an SO235 (UHF) connector and

attached the centre conductor to a piece of stiff wire running down the centre of the tin. I then soldered the resistors between this central conductor and the body of the tin. In this way the tin also served as a heat sink for the resistors as well as a shield for the dummy load.

6.4 The Voltage Divider

Two resistors in series can be used as a *voltage divider*. Consider the circuit below:



Voltage divider

This diagram shows two resistors connected in series as before. However, this time, we are measuring the voltage V_{OUT} across one of the resistors. Our task is to find this output voltage in terms of the input voltage applied across both resistors.

Using our formula for resistors in series, we know that the total combined resistance of R_1 and R_2 in series is $R_1 + R_2$. We can apply Ohm's Law to the input voltage and the combined resistance of R_1 and R_2 in series to find the input current I:

$$I = V_{In} / (R_1 + R_2)$$

If we assume that negligible current is drawn from the output, the same current I flows through both resistors. Hence we can find the voltage across R_2, which is the output voltage, using Ohm's Law:

$$V_{Out} = I R_2$$

Substituting the value we obtained for I by applying Ohm's Law to the series combination of R_1 and R_2 we get

$$V_{Out} = (V_{In} / (R_1 + R_2)) R_2$$
$$= V_{In} R_2 / (R_1 + R_2)$$

The circuit is known as a "voltage divider" because the output voltage is proportional to but smaller than the input voltage, so the effect of the circuit is to divide the input voltage by a constant (greater than 1).

Summary

Kirchoff's Current Law states that any point in a circuit where two or more conductors are joined, the sum of the currents flowing into the point is equal to the sum of the currents flowing away from the point. His Voltage Law states that the voltage between any two points in a circuit is equal to the sum of the voltage drops along any path connecting those points.

We can use these laws in conjunction with Ohm's Law to calculate the equivalent values of resistors in series and in parallel (or shunt). When two or more resistors are connected in series, the combined resistance is the sum of the individual resistances. When two or more resistors are connected in parallel, the reciprocal of the equivalent parallel resistance is the sum of the reciprocals of the individual resistances.

The voltage divider consists of two resistors in series with an output voltage measured across one of the resistors. The formula for the output voltage of a voltage divider is:

$$V_{Out} = V_{In}\ R_2 / (R_1 + R_2)$$

Revision Questions

1 Two 10 kΩ resistors are connected in parallel. If the voltage from a battery across the resistors sets up a current of 5 mA in the one resistor, how much current flows in the second resistor?

a. 10 mA
b. 2 mA
c. 20 mA
d. 5 mA

2 Two resistors are connected in series to a 9 V battery. The voltage across one of the resistors is 5 V. What is the voltage across the other resistor?

a. 4 V
b. 5 V
c. 9 V
d. 13 V

3 In a parallel circuit with a voltage source and several branch resistors, what relationship does the total current have to the current in the branch currents?

a. The total equals the average of the branch current in each resistor.
b. The total equals the sum of the branch currents in each resistor.
c. The total decreases as more parallel resistors are added to the circuit.
d. The total is calculated by adding the voltage drops across each resistor and multiplying the sum by the total number of all circuit resistors.

4 Two resistors are connected in series. The combined resistance is 1200 Ω. If one of the resistors is 800 Ω, what is the value of the other?

a. 1000 Ω
b. 800 Ω
c. 400 Ω
d. 1200 Ω

5 A 100 Ω resistor is connected in series with a 200 Ω resistor. The equivalent resistance of the two resistors is:

a. 100 Ω
b. 200 Ω
c. 300 Ω
d. 400 Ω

6 A 100 Ω resistor is connected in parallel with a 200 Ω resistor. The equivalent resistance of the two resistors is:

a. 50 Ω
b. 67 Ω
c. 75 Ω
d. 300 Ω

7 Three light bulbs are connected in series. Which of the following statements is necessarily true?

a. The current flowing through each of the bulbs is identical.
b. The voltage across each of the bulbs is identical.
c. The resistance of each of the bulbs is identical.
d. The light given off by each of the bulbs is identical.

8 Two light bulbs are connected in parallel to the mains supply. One of them blows, and becomes an open circuit (i.e. no current can flow through it). What will happen to the current flowing through the bulb that is still working?

a. Twice the current as before will flow through the working bulb.
b. No current will flow through the working bulb.
c. The same current as before will flow through the working bulb.
d. Half the current as before will flow through the working bulb.

9 The output voltage from a voltage divider with two equal resistances will be:

a. The same as the input voltage.
b. One quarter of the input voltage.
c. Half the input voltage.
d. One third of the input voltage.

10 A dummy load is made by connecting forty-four 2K2 resistors in parallel. The resistance of the dummy load is:

a. 20 Ω
b. 50 Ω
c. 75 Ω
d. 100 Ω

Chapter 7: Power in DC Circuits

7.1 Power Dissipation in Resistances

When a current flows through a resistance, the resistance will dissipate ("use up") power and generate heat. This principle is used in many electrical devices, for instance in electric bar heaters and kettles, where the elements are just resistances with suitable power handling and heat transfer abilities.

To calculate the power dissipated by a resistance, multiply the voltage across the resistance by the current flowing through the resistance:

$$P = V I$$

It is easy to see why. Remember that the voltage between two points is the amount of energy that it would take to move one unit of charge from the point of lower potential to the point of higher potential. Now that we are allowing the charge to flow from the point of higher potential back to the point of lower potential, this energy is recovered, usually in the form of heat. Since current is the rate of flow of charge, the greater the current, the greater the energy that will be given off each second. The rate at which energy is used is known as *power*.

For example, suppose an electric kettle draws 5 A at 240 V. Its power consumption is calculated as follows:

$$\begin{aligned} P &= V I \\ &= 240\ V \times 5\ A \\ &= 1200\ W \\ &= 1{,}2\ kW \end{aligned}$$

Although kettles usually work off AC not DC power, the section on AC power will show that the same formula applies.

7.2 Using Ohm's Law with the Formula for Power

Ohm's Law also deals with voltages and currents (as well as resistances), so it can often be used together with the formula for power. For example, suppose that in the example above we had instead been told that the kettle runs off 240 V and its element has a resistance of 48 Ω. We could then use Ohm's Law to calculate the current, since

$$\begin{aligned} I &= V \div R \\ &= 240\ V \div 48\ \Omega \\ &= 5\ A \end{aligned}$$

The rest of the calculation would then proceed as above, giving us the same answer of 1,1 kW. Another way is to combine Ohm's Law and the formula for power dissipation first, and only bring the actual numbers in at the end.

The formula for power is

$$P = V I$$

But according to Ohm's Law, we also know that

$$I = V \div R$$

So we can replace the symbol "*I*" in the power equation with "*V / R*" to give

$$P = V V \div R$$

And since *V x V* is just V^2 (pronounced "V squared"), we end up with

$$P = V^2 \div R$$

Applying this to the example, where V = 240 V and R is 48 Ω, we get

$$\begin{aligned} P &= (240\ V)^2 \div 48\ \Omega \\ &= 57\,600\ V^2 \div 48\ \Omega \\ &= 1200\ W \\ &= 1{,}2\ kW \end{aligned}$$

Which fortunately is the same answer as before.

In the same way, if you know the current flowing through a resistance and the value of the resistance, but not the voltage across it, you can use Ohm's Law to calculate the voltage across the resistance and then apply the formula for power to calculate the power dissipation. Or these two steps can be combined in a single equation:

	P	$= V I$	(the formula for power)
and	V	$= I R$	(Ohm's Law)
so	P	$= I I R$	
		$= I^2 R$	

We now have a simple formula for calculating power from current and resistance:

$$P = I^2 R$$

For example, suppose a 50 Ω resistor has a current of 2 A flowing through it. The power dissipated is:

$$\begin{aligned} P &= I^2 R \\ &= (2A)^2 \times 50\ \Omega \\ &= 4\ A^2 \times 50\ \Omega \\ &= 200\ W \end{aligned}$$

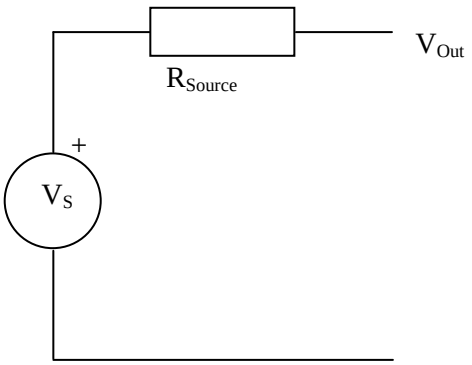
Exercise

Use Ohm's Law to find the voltage across the resistor, and then the formula $P = V I$ to calculate the power dissipated by the resistor, and see if you get the same answer.

7.3 Electrical Sources

Real-life electrical sources, including batteries and mains-powered supplies, do not always behave in the manner we would prefer. Normally, the voltage drops when we start to draw current from the source.

A real-life supply can be modelled in two ways. The simplest to work with in most cases is a constant-voltage supply with some internal resistance. This model is known as a *Norton equivalent*.

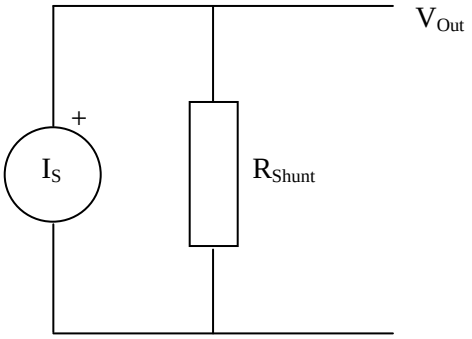


Constant-voltage supply with internal resistance

The circle with V_S in it represents a perfect voltage supply, which maintains a constant voltage of V_S. The internal resistance is represented by R_{Source}.

In this case, $V_{Out} = V_S$ when no current is flowing. When current is drawn from the power supply, a voltage $V_{Drop} = I \times R_{Source}$ develops across the internal resistance. The output voltage then decreases to $V_{Out} = V_{Source} - V_{Drop}$. Obviously, the lower the internal resistance, the smaller the voltage drop will be, and the more constant the output voltage will be. Well-regulated power supplies and good-quality batteries have low internal resistance.

Another way to model the same power supply is by a constant-current source with a shunt resistance. This model is known as a *Thevenin equivalent*.



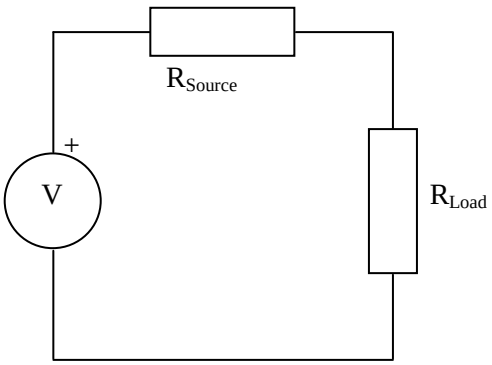
Constant-current supply with shunt resistance

If the model is a good one, the behaviour must be exactly the same as the previous model's behaviour. If no current is being drawn, $V_{Out} = I_S \times R_{Shunt}$, as that voltage is developed across R_{Shunt} due to the current. As current is drawn by a load, less current flows in R_{Shunt}, as the total current remains constant, but is divided between R_{Shunt} and the load. In this case, V_{Out} decreases as the load current increases. It seems like a suitable choice of I_S and R_{Shunt} will result in the same behaviour we observed previously.

For the remainder of this course, we will stick to constant-voltage sources with series resistance, as the Norton model is somewhat easier to visualise.

7.4 Matching Source and Load

A practical power supply with some internal resistance and a load is represented in the diagram below. R_{Load} is the load resistance. A *load* is something to which the circuit is delivering power. Depending on the application it might be an antenna, an electric motor, a light bulb or anything else that uses power.



Constant-voltage supply with internal and load resistors

An interesting question is: What load resistance (i.e. what value of R_{Load}) will result in the maximum power transfer to the load?

If the load resistance is very low, a lot of current will flow in the circuit, but the voltage across the load will be small. If the resistance is high, the voltage across the load will be high, but the current through it will be low. Since $P = V I$, both the current through the load and the voltage across it are important for power transfer.

Although the mathematics is a bit beyond the level of this course, it turns out that the load dissipates the maximum power when the load resistance is exactly equal to the source resistance. In this case, the power dissipated by the load is $V^2 \div (4 \times R_{load})$. This result is useful to know when designing power sources such as power amplifiers. You should note, however, that with a matched load the source dissipates just as much power as the load, so cooling may be quite important! Cooling could be achieved using fans or *heat sinks*, finned metal objects that allow heat to be safely exchanged to the air. The fins help to increase the contact area with the air, for optimal heat transfer. In some high-power applications, liquid cooling may even be used.

Summary

The power dissipated in a resistive load can be found using the formula $P = V I$. This formula can be combined with Ohm's Law to give $P = I^2 R$ and $P = V^2 / R$. In a simple resistor, this power will be dissipated as heat.

All real-life voltage sources have some internal resistance. The internal resistance causes the output voltage to drop somewhat when current is drawn. The maximum power transfer from the source to the load occurs when the load resistance is exactly equal to the source resistance.

Revision Questions

1 A light bulb is rated at 12 V and 3 W. The current drawn when used with a 12 V source is:

a. 250 mA
b. 750 mA
c. 4 A
d. 36 A

2 The DC current drawn by the final stage of a linear amplifier is 100 mA at 100 V. How much power is consumed?

a. 100 W
b. 1 kW
c. 10 W
d. 1 W

3 If a power supply delivers 200 W of electrical power at 400 V DC to a load, how much current does the load draw?

a. 0,5 A
b. 2 A
c. 5 A
d. 80 kA

4 The product of the current and what force gives you the electrical power in a circuit?

a. Magnetomotive force.
b. Centripetal force.
c. Electrochemical force.
d. Electromotive force.

5 A resistor is rated at 10 W. Which of the following combinations of potential difference and current exceeds the rating of the resistor?

a. 2 V at 100 mA
b. 20 V at 200 μA
c. 1 kV at 25 mA
d. 0 mV at 2 A

6 The starter motor of a motor car draws 200 A from the 12 V battery. How much power does it use?

a. 24 W
a. 240 W
b. 2,4 kW
c. 24 kW

7 What is the resistance of the motor in question 6?

a. 60 mΩ
b. 100 mΩ
c. 600 mΩ
d. 1 Ω

8 The internal resistance of a car battery is found to be 0,2 Ω. Into what load resistance will it deliver the maximum power?

a. 0,1 Ω
b. 0,2 Ω
c. 0,6 Ω
d. 1,2 Ω

9 At its peak, a lightning bolt has a voltage of 100 MV and a current of 10 kA. How much power does it deliver at that moment?

a. 10^9 W
b. 10^{10} W
c. 10^{11} W
d. 10^{12} W

10 A current of 2 mA is measured in a 1 kΩ resistor. How much power is the resistor dissipating?

a. 2 mW
b. 4 mW
c. 2 W
d. 4 W

Chapter 8: Alternating Current

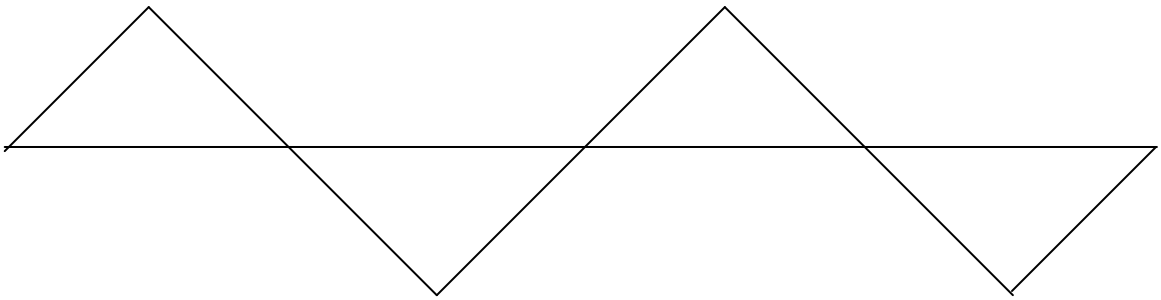
8.1 Introduction

In direct current (DC) circuits, the current always flows in one direction. The two terminals of the voltage sources used to power these circuits always have the same polarity – one terminal is always positive with respect to the other terminal. This positive voltage causes the current to flow in only one direction in the circuit.

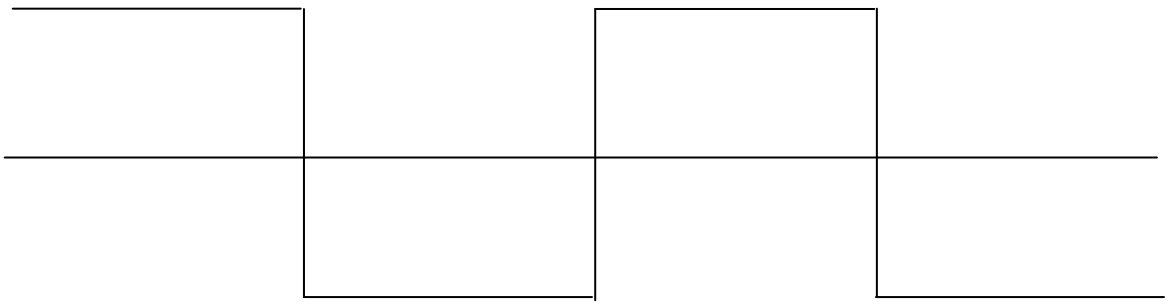
In some circuits, though, the direction in which the current flows is constantly changing. The current flows first in one direction, then in the reverse direction, then in the original direction again and so on, with the direction changing at regular intervals, usually many times each second. The circuits are called *alternating current* (AC) circuits. Power for these circuits may be supplied by alternating current (AC) power supplies, such as the mains supply. With AC power supplies, there is no "positive" or "negative" terminal. Instead, one terminal will be positive with respect to the other for a brief period, and then the roles will reverse and the other terminal will become more positive for a brief period, and so on. Although the abbreviation AC stands for "alternating current", it is also used to refer to voltages, in phrases such as "An AC Voltage" and "15 V AC".

8.2 The Sine Signal

If you were to plot the voltage or current in an AC circuit against time, there are many possible shapes (known as "waveforms") that they could take. For example:



A Triangular Signal



A Square Signal

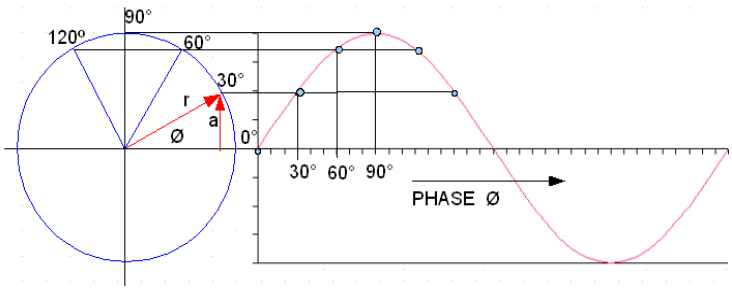
However, when we analyze AC circuits, we normally think of the waveform as being a "sine signal". A sine function is a mathematical concept, but it describes something that occurs frequently in nature. Mechanical vibrations, rotating bodies and electrical circuits, among others, can be described with the same mathematical models.

To understand what a sine function is, we start by looking at rotating bodies. Visualise a point on the circumference of a rotating wheel. Now think only about the vertical

displacement of the point from the central axis, the quantity marked as *a* in the figure below. If the wheel rotates at a constant speed, a sine function is produced by the value of *a*. To the right of the wheel is a graph. The horizontal axis represents the phase or angular displacement of the wheel from some arbitrary starting point, starting at 0°. The amplitude above (or below) the zero line is transferred to the graph as well. The relationship between the angle Ø, the radius of the circle, *r*, and the amplitude *a* is given by:

$$\sin Ø = a \div r$$

The amplitude is then given by $a = r \sin Ø$ and hence the term "sine function".



Generation of a sine signal

Note that the amplitude values start decreasing after 90°. The amplitude at 120° is the same as that for 60°. After rotating through half of a revolution, the values become negative (below the line). The shape is the same, but inverted.

Each rotation of the wheel represents one cycle of the waveform, from an angle of 0° to an angle of 360°.

In electrical circuits, the sine function is applied as:

$$V = V_{Peak} \sin (2\pi ft)$$

where V_{Peak} is the peak voltage of the waveform, f is its frequency, t is time, π is the mathematical constant "pi" (approximately 3,14) and *sin* is the trigonometric sine function. As time increases, the sine function goes from 0 to 1 to 0 to -1 to 0 to 1 and so on. At the same time, V goes from 0 to V_{Peak} to 0 to - V_{Peak} to 0 to V_{Peak} and so on.

Note that the sine function does *not* consist of two semi-circles, which is how it is sometimes incorrectly drawn.

The reason why we deal mostly with sine functions in circuit analysis is because the French mathematician Joseph Fourier (1768-1830) showed that any other periodic function could be decomposed into a number of sine functions of different frequencies. So if we know how a circuit responds to a sine signal, we can easily calculate its response to any other signal using the technique known as *Fourier analysis*. A sine signal represents a "pure" AC signal that contains only a single frequency, known as the *fundamental*. Any other signal includes both the fundamental and *harmonics*, which are integral multiples of the fundamental frequency.

You will often hear the term "*sine wave*". Although most radio signals are indeed sine waves as they travel through space, we are more interested in sine-shaped signals in electric circuits. To avoid clumsy constructions like *sinusoidal signal*, we therefore use the term "*sine signal*" to describe the voltage and current shapes inside electrical circuits.

8.3 Cycles and Half Cycles

An AC signal consists of many identical *cycles* one after another. The figure above shows one complete cycle of a sine signal, while above it are two complete cycles of square and triangular signals.

Question: *How many cycles of a square signal are shown in the first figure of this section?*

Usually electrical waveforms like AC voltages and currents are positive for half the time and negative for the other half. When we want to refer just to the positive or negative period, we speak of the "*positive half cycle*" and "*negative half cycle*".

8.4 Period and Frequency

The period of a waveform is the time taken for one cycle, which is usually expressed in seconds, milliseconds or microseconds (s, ms or µs).

Definition: *The period of a waveform is the time taken for one complete cycle.*

A complete cycle can start at any point on the waveform, and continues to the corresponding point on the next cycle. In each of the preceding figures, the cycle starts at 0 and ends at 0. However, the cycle could equally well have started at the maximum (1) or at any other point (e.g. when the value is 0,462).

The frequency of a waveform is the number of cycles per second. The unit of frequency, one cycle per second, is called the hertz (abbreviated Hz) in honour of the German physicist Heinrich Hertz (1857-1894). The dimension of the unit Hz is the same as "per second".

Definition: *The frequency of a waveform is the number of cycles per second.*

Since period is the number of seconds per cycle, and frequency is the number of cycles per second, it follows that the period and frequency of a waveform are reciprocals of each other. That is:

$$t = 1 \div f$$

and

$$f = 1 \div t$$

where t is the period (in s) and f the frequency (in Hz).

For example, the mains frequency in South Africa is 50 Hz (50 cycles per second). The period can be found from:

$$t = 1 \div f$$
$$= 1 \div 50\ Hz$$
$$= 0{,}02\ s$$
$$= 20\ ms$$

8.5 Wavelength and the Speed of Light

Electrical currents and voltages move through wires at almost the speed of light, which is a very high but not infinite speed. Radio waves transmitted from an antenna also travel at the speed of light. The speed of light, which is usually represented by the symbol c in physics, is approximately 3×10^8 m/s, equivalent to just over 1 000 000 000 km/h!

Think about an AC waveform with a constant frequency moving through a very long wire at the speed of light. The start of one cycle will occur at a particular point in time, and hence at a particular point along the wire. The start of the next cycle will occur a certain time later (this time difference being the *period* of the wave), during which the wave, which

is traveling at the speed of light, will have moved some distance further along the wire. Since the speed of light is constant, and the time between successive cycles of the wave (the period) is also constant, the distance traveled by the wave between the start of one cycle and the start of the next must also be constant for this particular wave. This distance is known as the *wavelength* of the wave.

Definition: *The wavelength of a wave is the distance it travels in one cycle.*

Because wavelength represents the distance the wave travels during a certain time, it is related to the period and frequency as follows:

$$\lambda = c\,t$$
and $$\lambda = c / f$$

where λ (a Greek lower case "L", pronounced "lambda") is the wavelength in m, *c* the speed of light in m/s, *t* the period in s and *f* the frequency in Hertz.

For example, one of the author's favourite radio stations is Cape Talk, which broadcasts on a frequency of 567 kHz. The corresponding wavelength can be calculated as follows:

$$\lambda = c \div f$$
$$= 3 \times 10^8 \text{ m/s} \div 567 \times 10^3 \text{ Hz}$$
$$= 529 \text{ m}$$

This is the distance that the radio waves transmitted by Cape Talk will travel during one complete cycle.

There is a short cut that is quite useful for radio amateurs. Because we express most of our frequencies in megahertz (millions of cycles per second), you can avoid having to deal with lots of zeros (or with scientific notation) by using the formula:

$$\lambda = 300 \div F$$

where λ is the wavelength in m and *F* the frequency in MHz. For example, a frequency of 14,100 MHz has a wavelength of:

$$\lambda = 300 \div F$$
$$= 300 \div 14{,}1 \text{ m}$$
$$= 21{,}3 \text{ m}$$

Note that the higher the frequency, the shorter the wavelength and *vice versa*. You can also calculate the frequency from the wavelength using the formula:

$$F = 300 \div \lambda$$

where F is the frequency in MHz and λ the wavelength in m. For example, the amateur "two-meter" band has frequencies of approximately:

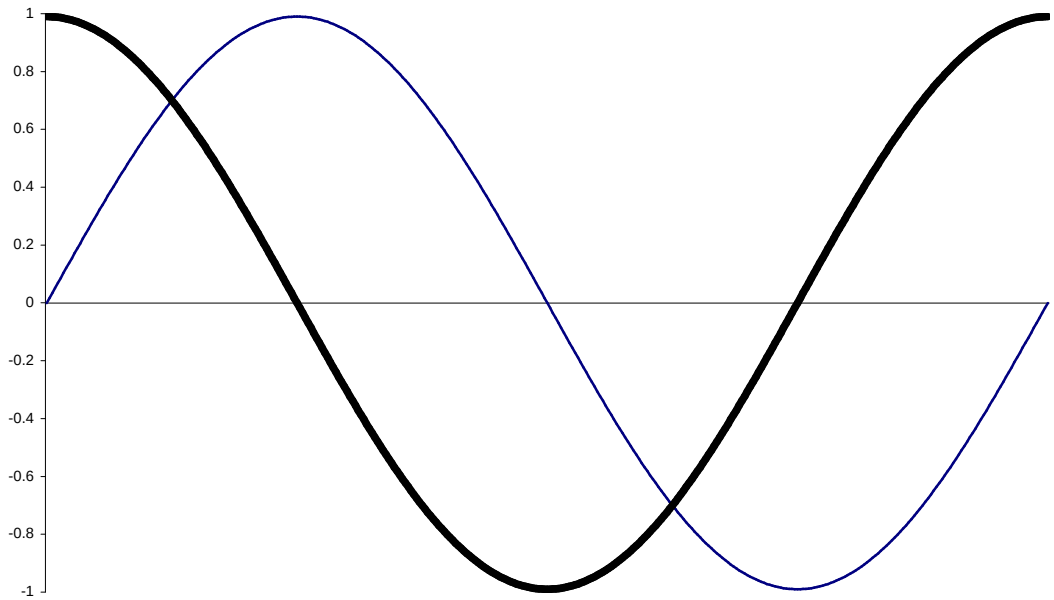
$$F = 300 \div \lambda$$
$$= 300 \div 2 \text{ MHz}$$
$$= 150 \text{ MHz}$$

The actual frequency limits of the two-meter band in South Africa are 144 and 146 MHz. The reason for the discrepancy is that "two-meter band" is intended as a name for the band,

not an accurate representation of its wavelength. “Two comma zero seven meter band” would be a bit of a mouthful!

8.6 Phase

It is possible to have two sine signals of the same frequency but where the cycles start at different times. In this case, we talk of the waves having a *phase difference*. The phase difference is usually expressed in degrees. One complete cycle comprises 360º so for example a phase difference of one quarter of a cycle would be 90º.



Two sine signals with a phase difference of 90º

The wave that reaches a certain part of its cycle before the other is said to *lead* the other wave. Conversely, the wave that reaches that part of its cycle after the other wave is said to *lag* the other wave. Any point on the cycle could be used, as long as corresponding points are used for both signals.

In the figure above, the wave drawn with a thick line *leads* the other wave by 90º because it gets to each value *before* the other wave does. It leads by 90° at any arbitrarily chosen point. In this case, the maximum (1) or the zero crossing (0) is easy to see.

8.7 RMS Voltage and Current

Remember the formulae to find power dissipation given the value of a resistance and the voltage across the resistance:

$$P = V^2 \div R$$

If this formula is applied to sine signal, one can see that the power dissipation is at a maximum at the positive and negative peaks of the signal, and at a minimum when the voltage is zero. Remember that the V^2 is positive even when the voltage is negative; the square of a negative number is a positive number.

If we were able to average out the square of the voltage through a full cycle of the sine signal, we could calculate an equivalent DC voltage that would cause the same power dissipation in a resistor. This voltage is known as the “root mean square” or RMS voltage.

In this context, "Mean" indicates the average, so RMS means "the square root of the average of the square of all the voltages in a cycle".

> **Definition:** *The RMS value of an AC voltage is the value of the DC voltage that would cause the same power dissipation in a resistance.*

For a sine signal, the RMS voltage is the peak voltage (the maximum voltage reached on both positive and negative peaks) divided by the square root of two.

$$V_{RMS} = V_{Peak} \div \sqrt{2}$$
so $$V_{RMS} = 0{,}707\ V_{Peak}$$

Note that this formula only works for a sine signal, as the ratio may be different for other waveforms.

Whenever one gives the value of an AC voltage, the value is assumed to be the RMS value unless specifically otherwise noted. For example, the mains voltage in South Africa is specified as 240 V AC. This is the *RMS* value. We can calculate the peak value as follows:

$$V_{RMS} = V_{Peak} \div \sqrt{2}$$
so $$V_{Peak} = \sqrt{2}\ V_{RMS}$$
$$= 1{,}41 \times 240\ V$$
$$= 338\ V$$

In the same way, AC current is usually expressed as an RMS current unless otherwise specified. The definition is similar:

Definition: *The RMS value of an AC current is the value of the DC current that would cause the same power dissipation in a resistance.*

The RMS current can be found from the peak current using a similar formula to the one used to find the RMS voltage from the peak voltage:

$$I_{RMS} = I_{Peak} \div \sqrt{2}$$
so $$I_{RMS} = 0{,}707\ I_{Peak}$$

The nice thing about working with RMS voltages and currents is that Ohm's Law and the formulae for power work for AC voltages and currents just like they do for DC voltages and currents, as long as you use the RMS values.

For example, if the element of a kettle that runs of 240 V AC (RMS) has a resistance of 48 Ω, the current flowing through the element is

$$I = V \div R$$
$$= 240\ V \div 48\ \Omega$$
$$= 5\ A\ (RMS)$$

Although I have noted that the 5 A is an RMS value, this would not normally be necessary as RMS measurements are assumed for all AC values unless otherwise specified.

Similarly the power can be calculated using the usual formula,

$$P = V\ I$$
$$= 240\ V \times 5\ A$$
$$= 1{,}2\ kW$$

Because we are using RMS values, the standard formula gives the right answer.

8.8 Frequency Ranges

Because we are primarily interest in radio, we need to understand something about frequency ranges.

The terminology in the table below may seem somewhat strange. Remember that the technology required to operate on higher and higher frequencies took time to evolve. When this table was designed, "high" wasn't as high as it would be today, but the terms have been retained for historical reasons.

Range		Frequency		Wavelength		Amateur bands
		From	To	From	To	
LF	Low frequency	30 kHz	300 kHz	10 km	1 km	Experimental
MF	Medium frequency	300 kHz	3 MHz	1 km	100 m	160 m
HF	High frequency	3 MHz	30 MHz	100 m	10 m	80 to 10 m (9 bands)
VHF	Very high frequency	30 MHz	300 MHz	10 m	1 m	6 m, 2 m
UHF	Ultra-high frequency	300 MHz	3 GHz	1 m	100 mm	70 to 23 cm
SHF	Super-high frequency	3 GHz	30 GHz	100 mm	10 mm	Several
EHF	Extremely high frequency	30 GHz	300 GHz	10 mm	1 mm	Several
THF	Tremendously high frequency	300 GHz	3 THz	1 mm	100 µm	Several

The majority of amateurs operate on HF, VHF and perhaps UHF, so these are the ranges you must be familiar with.

Summary

AC signals consist of many identical cycles, one after another. Sine signals consist of a single frequency known as the *fundamental*. All other signals have additional frequencies, the *harmonics.* The period of a waveform is the time taken for one complete cycle. The frequency of a waveform is the number of cycles per second.

The wavelength of a wave is the distance it travels in one cycle. The wavelength and frequency of a wave are related by the formula:

$$F = 300 \div \lambda$$

where F is the frequency in MHz and λ the wavelength in m. Phase differences are expressed in degrees, with 360º in one complete cycle.

AC voltages and currents are expressed as RMS values. The RMS value of an AC voltage or current is the value of the DC voltage or current that would cause the same power dissipation in a resistance. For sine signals, the RMS voltage can be calculated from the peak voltage using the formula

$$V_{RMS} = 0{,}707\ V_{Peak}$$

For radio purposes, the frequency of AC current is divided into bands. The most important ones are:

Range		Frequency		Wavelength		Amateur bands
		From	To	From	To	
HF	High frequency	3 MHz	30 MHz	100 m	10 m	80 to 10 m (9 bands)
VHF	Very high frequency	30 MHz	300 MHz	10 m	1 m	6 m, 2 m
UHF	Ultra-high frequency	300 MHz	3 GHz	1 m	100 mm	70 to 23 cm

Revision Questions

1 The frequency of an AC waveform is defined in the unit:

a. seconds
b. velocity
c. period
d. hertz

2 The frequency of 5 Hz has a period of:

a. 2 s
b. 300 s
c. 200 ms
d. 1,2 s

3 The wavelength of a signal of 100 MHz in free space is:

a. 30 mm
b. 0,3 m
c. 3,0 m
d. 30 m

4 A radio wave has a period of 20 ms. Its wavelength in free space is:

a. 6 km
b. 60 km
c. 600 km
d. 6000 km

5 Two sine signals are 180º out of phase. When the one signal is at its maximum positive value the other:

a. Is also at its maximum positive value.
b. Is at its most negative value.
c. Is at zero.
d. Cannot be determined from the information given.

6 Which signal consists of only the fundamental frequency, without any harmonics?

a. A square signal.
b. A sine signal.
c. A triangular signal.
d. A saw-tooth signal.

7 Which value represents the ratio of RMS to Peak value of a sine AC waveform?

a. 0,5
b. 0,636
c. 1,414
d. 0,707

8 What is the value of an AC waveform, representing the equivalent heating effect to a DC voltage, known as?

a. RMS value.
b. Average value.
c. Peak value.
d. Corrected value.

7. The mains voltage in the U.S.A. is 115 V RMS. What is the peak voltage?

a. 81 V
b. 115 V
c. 163 V
d. 220 V

8. The mains voltage in South Africa is 240 V RMS. If this voltage is applied across a heating element with a resistance of 576 Ω, how much power will be dissipated?

a. 10 W
b. 57,6 W
c. 100 W
d. 576 W

11. An electric geyser operating from the 240 V AC RMS mains supply consumes 2,4 kW. What current does it draw?

a. 10 A RMS.
b. 10 A peak
c. 10 A average
d. 10 A DC

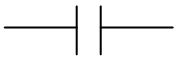
12. A hi-fi loudspeaker has a resistance of 8 Ω. When it is delivering 8 W, what is the RMS voltage across the speaker?

a. 1 V
b. 8 V
c. 10 V
d. 80 V

Chapter 9: Capacitance and the Capacitor

9.1 The Capacitor

The capacitor is a component that consists of two conductors in close proximity, separated by an insulating material known as the *dielectric*. The circuit symbol for a capacitor is quite suggestive of its construction:

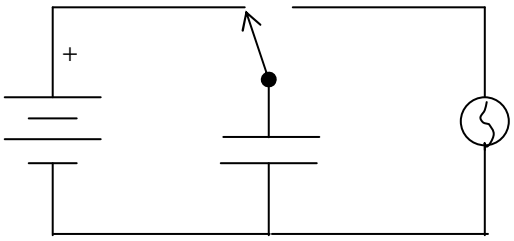


Symbol for capacitor

The two vertical lines represent the conductors and the gap between them represents the insulating dielectric.

Real capacitors may consist of plates or foil or blocks, and may be separated by air, plastic or liquid. Conceptually, we can think of two parallel plates separated by air. This arrangement is typical of variable capacitors, which we will discuss later in this section.

Capacitors have a property known as *capacitance*, which is the ability to store energy in an electric field between the plates. To see how this works, consider the circuit below:



Capacitor with battery and globe

This shows a capacitor connected to a switch that can either be used to connect it to the battery on the left or to the light bulb on the right.

Let us start by thinking what happens in terms of electrons. When the capacitor is connected to the battery, some of the negatively charged electrons in the upper plate of the capacitor are attracted towards the positive terminal of the battery, through the conductor between them. At the same time, some electrons flow from the negative terminal of the battery to the lower plate of the capacitor. In effect the battery is acting as an “electron pump”, pumping some electrons from the top plate of the capacitor, through the battery and to the bottom plate of the capacitor.

Through this process the upper plate of the capacitor loses some of its electrons, so it becomes positively charged. At the same time the lower plate gains some excess electrons so it becomes negatively charged. In terms of conventional current, a current flows through the capacitor from top to bottom, which generates a potential difference across the capacitor, with the upper plate becoming more positively charged and the lower plate becoming more negatively charged. This process is known as “charging” the capacitor.

The voltage that is developed across the capacitor opposes the flow of current through the capacitor. As the voltage across the capacitor increases the current through it decreases and when the voltage across the capacitor is equal to the battery voltage the current stops flowing altogether. The capacitor is now fully charged.

Assume that the switch is now flipped so that the capacitor is connected to the light bulb. The excess of electrons on the negatively charged lower plate will be attracted to the positively charged upper plate, which has a shortage of electrons, so a current will flow. In conventional terms, the current flows from the positively charged upper plate to the negatively charged lower plate. This current will make the light bulb glow. As the current flows, the charges on the capacitor plates will gradually return to neutral, and the voltage across the plates will reduce. This reduced voltage will reduce the current flowing in the circuit until eventually both plates have the same concentration of electrons. There is now no potential difference across the capacitor and the current will stop flowing altogether. This process is known as "discharging" the capacitor.

When a capacitor is charged in a DC circuit it has a voltage across it and a current flowing through it, so power is being provided to the capacitor according to the formula $P = V I$. However, this power is not being dissipated into heat as it was in a resistor. Instead, energy is being stored in the electric field between the plates. When the capacitor is discharged this energy is released – in this case, it causes the light bulb to glow.

Capacitors come in different values. The value of a capacitor (its capacitance) depends on:

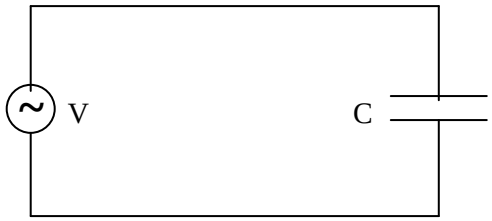
- **The surface area of the plates.** The greater the area, the greater the capacitance.
- **The distance between the plates.** The greater the distance, the lower the capacitance.
- **The dielectric constant.** This constant is property of the dielectric. The larger the dielectric constant, the greater the capacitance. The dielectric constant of a vacuum is 1, with air being very close to 1.

Large capacitors (meaning those with high capacitance, not necessarily related to their physical size) are able to store a lot of energy by allowing a large excess of positive or negative charge to accumulate on the plates. Small capacitors can only store a little energy, as only a small amount of excess charge can be accumulated. The value of a capacitor is measured in farad (abbreviation F), and typical practical capacitors range in size from 1 pF to 1000 μF or so. Some capacitors are marked with colour codes similar to resistor codes, while others are marked with stenciled lettering. Because the "μ" in "μF" appeared to cause problems for manufacturers, the abbreviation "MFD" is often marked on capacitors.

Extending our previous flow analogies, a capacitor can be modelled as a diaphragm, which can store energy by flexing if a pressure difference is applied across it. Once the pressure is removed, the fluid will flow from the pressurised end of the diaphragm and into the unpressurised end, until the pressure difference is relieved. Note that no water flows through the diaphragm, just like no electrons actually flow through the dielectric.

9.2 Capacitors in AC Circuits

Capacitors get more interesting in AC circuits. Consider this circuit, which shows an AC voltage source connected to a capacitor through a resistor:



AC voltage source and capacitor

The "~" symbol on the voltage source means that it is an AC source. The symbol "V" represents the voltage of the source, and "C" represents the value of the capacitor.

The first question is whether a current will flow at all. If the voltage source was DC then the capacitor would soon charge up to the same voltage as the voltage source, and no more current would flow, except possibly for a very small *leakage* current. However, because in this circuit we have an AC voltage source, the situation is different. As the current flows in one direction, the capacitor will begin to charge up and the potential difference this causes will oppose the flow of current through the capacitor. However, when the current changes direction, the capacitor will start to discharge and the energy it had "borrowed" will be returned to the circuit. Eventually the capacitor will be fully discharged and will start to charge up again but with the reverse polarity. Then when the current direction reverses again, the capacitor can discharge again before once again charging in the original direction.

With an AC source, an AC current will flow through a capacitor. It is interesting to consider the effect of frequency. A low frequency AC source will cause the current to flow for a long time in one direction. During this time, the capacitor will become appreciably charged and the potential difference that forms across its plates will significantly oppose the flow of current in the circuit. For a low frequency source, therefore, only a small current will flow. On the other hand, with a high frequency source current will only flow in one direction for a short time before reversing direction. This delay will not be long enough to charge the capacitor much, so not much potential difference will develop across the plates, and there will not be much opposition to the flow of current. For a high frequency source, therefore, a larger current will flow.

9.3 Capacitive Reactance

The opposition to the flow of current that we experience with capacitors in an AC circuit is not resistance. If it were resistance, power would be dissipated by the capacitor. However, we have seen that the energy that is "borrowed" during one half cycle is returned to the circuit during the next half cycle. The opposition to the flow of current in a capacitor is called "capacitive reactance" and usually given the symbol X_C. The formula for the reactance of a capacitor is:

$$X_C = 1 / (2 \pi f C)$$

where X_C is the capacitive reactance in Ω, f the frequency in Hz and C the capacitance in farad (abbreviation F). Note that the reactance *decreases* as the frequency *increases*. This trend is because capacitors oppose the flow of current less and less as frequency increases.

Our original question was how much current will flow in the circuit. Fortunately, Ohm's Law works for reactance in just the same way as it does for resistance:

$$I = V / X$$

Once we have calculated the reactance of the capacitor, we can easily calculate the current flowing in the circuit. However, note that although resistance and reactance are both measured in Ω, you cannot simply add them together. You will understand later how these quantities can be related.

For example, in the circuit above suppose the voltage V is 1 V, the capacitance of the capacitor is 1 nF (10^{-9} F) and the frequency is 1 MHz (10^{6} Hz), the reactance of the capacitor is:

$$X_C = 1 / (2 \pi f C)$$

$= 1 / (2 \times 3{,}14 \times 10^{6}\ Hz \times 10^{-9}\ F)$
$= 1 / (0{,}006\ 28\ F.Hz)$
$= 159\ \Omega$

The current flowing in the circuit can be found using Ohm's Law in a slightly modified form:

$$I = V / |X_C|$$

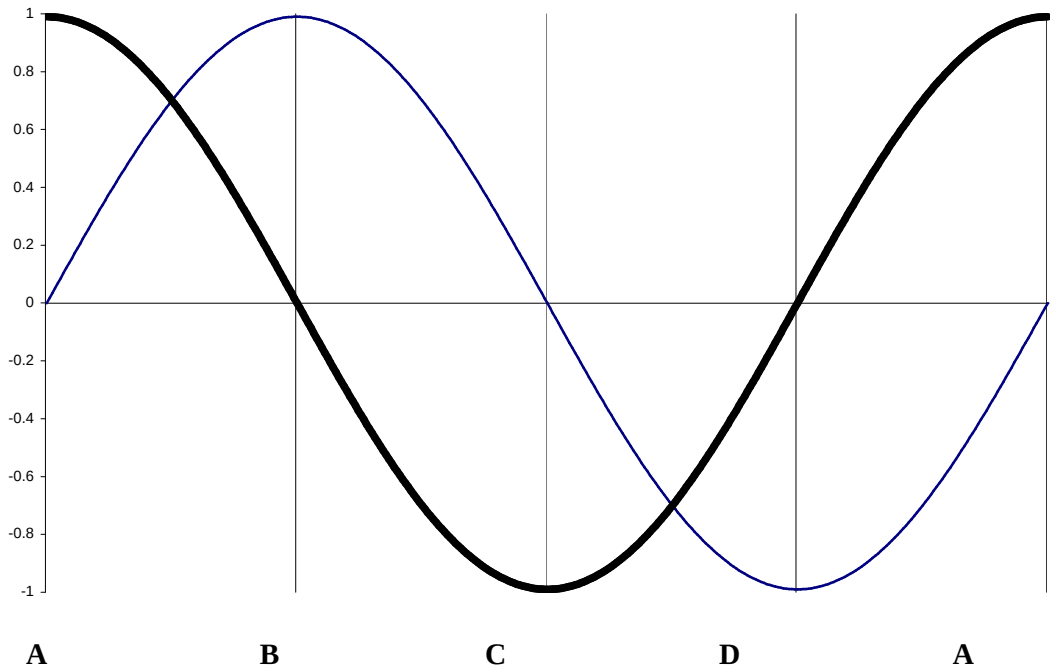
Here, |X| means "the magnitude of X", in other words the value of X but without the minus sign if it has one. So

$I = 1\ V / 159\ \Omega$
$= 0{,}0063\ A$
$= 6{,}3\ mA$

9.4 Phase of Current and Voltage

The current flowing through a capacitor and the voltage across the capacitor have an interesting property: they are always 90º out of phase. To be precise, the current flowing through a capacitor *leads* the voltage across the capacitor by 90º, so the voltage across the capacitor *lags* the current flowing through the capacitor by 90º. In the graph below, the thick line represents the current through the capacitor, while the thin line represents the voltage across the capacitor.

The material within this box is for developing insight and will not form part of the RAE.



This result is not as counterintuitive as it might seem at first glance. Remember that the capacitor is charging – that is, the voltage across its plates is increasing – for as long as there is current flowing through it in the right direction. The capacitor should reach maximum positive charge – i.e. with the maximum positive voltage across the plates – when a positive current has been flowing through it for as long as possible. Time B in the diagram above indicates this condition. Similarly, the capacitor should reach maximum negative charge at the point where a negative current has been flowing through it for as long as possible, which

it does at time D.

Also, since the rate at which a capacitor charges or discharges depends on the current flowing through it, this rate should be greatest at the points of maximum current. For example, at time A where the maximum positive current is flowing through the capacitor, the rate at which it is charging is greatest. The voltage therefore rises most rapidly at this point. Similarly at time C, where the maximum negative current flows through the capacitor, is where its rate of discharge is greatest, and the voltage drops most rapidly.

We mentioned before that reactance differs from resistance, even though they are both measured in Ω. The voltage across a resistance is always in phase with the current through the resistance, while the voltage across a reactance is always 90º out of phase to the current flowing through the reactance. In fact, this explains why there is no power dissipated by a reactance. The formula for power is:

$$P = V I$$

However, remember that a positive number multiplied by another positive number provides a positive result, as does a negative number multiplied by another negative number. A positive number multiplied by a negative number or *vice versa* produces a negative result.

If you look at the graph above showing the voltage across and current through a capacitor, you will see that in the first quarter of the graph from time A to time B, both voltage and current are positive, so the power "dissipated" is positive. However, in the last quarter of the graph, between time D and time A, the voltages and currents have exactly the same values (although in reverse), but this time the voltage is negative while the current remains positive, so the overall result is negative. This negative "dissipation" precisely cancels out the positive power dissipation in the first quarter of the graph.

Similarly, between C and D the voltage and current are both negative, so the result is a positive power "dissipation". However, the voltage and current have exactly the same values between B and C (again in reverse), but this time the voltage is positive while the current remains negative, so the overall power "dissipation" is negative and exactly cancels out the positive power dissipation between C and D.

So the positive dissipation from A to B and from C to D is exactly cancelled out by the negative "dissipation" from B to C and from D to A. The capacitor is "borrowing" energy as it charges, only to "return" it as it discharges.

9.5 Capacitors in Parallel and Series

Two or more capacitors connected in parallel are equivalent to a single capacitance with a value equal to the sum of the values of the individual capacitors.

So for capacitors connected in parallel,

$$C_{Total} = C_1 + C_2 + \ldots$$

Note that this equation is similar to the equation for resistors in *series*.

For capacitors connected in series,

$$1/C_{Total} = 1/C_1 + 1/C_2 + \ldots$$

Note that this result is similar to the equation for resistors in *parallel*.

9.6 Types of Capacitor

Many capacitors are not really a pair of plates separated by a dielectric. Like resistors, capacitors come in several different types that are designed for different applications.

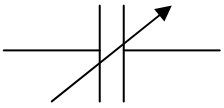
- **Ceramic capacitors** are generally good for radio frequency (RF) applications and are inexpensive but their tolerance is poor (around ±10%) so they should not be used in critical applications such as the frequency-determining elements in oscillators or filters. They are available in values ranging from 100 pF to 100 nF or so, and in high voltage ratings up to 15 kV.

- **Silvered Mica capacitors** are also good at RF and have much tighter tolerances (typically ±1%) but are quite expensive. They are only available in fairly small values from 1 pF to 100 nF.

- **Polycarbonate capacitors** are suitable when higher capacitance values are required at medium tolerances (±5% is typical). Values range from 10 nF to 10 μF.

- **Electrolytic capacitors** use metal (usually aluminium) foil as one "plate" of the capacitor and a conductive fluid as the other "plate". The insulating dielectric is a very thin chemical layer that is deposited on the metal film by the dielectric fluid. Electrolytic capacitors can have very high values, up to 100 F, but most of them are *polarised* meaning that one of the terminals must always be positive with respect to the other. This limitation makes them most suited to DC applications like power supplies.

- **Variable capacitors** consist of two sets of plates. Turning the control knob moves one of the sets of plates (the *rotor*) and varies how much they overlap the set of fixed plates (the *stator*). In this way the capacitance can be varied. Variable capacitors were once used as the tuning controls of radios. The symbol for a variable capacitor is shown below:



Symbol for variable capacitor

Summary

Capacitors consist of two conductors separated by an insulating dielectric. Capacitors have a property known as *capacitance*, which is the ability to store energy in an electric field between the conductors. The energy is stored when the capacitor is *charged* and released when it is *discharged*. The capacitance of a capacitor depends on the surface area of the conductors, the distance between the conductors and the dielectric constant of the insulating material.

In AC circuits, capacitors exhibit *reactance* which opposes the flow of current. Although reactance is measured in Ω, it is not the same as resistance as no energy is being dissipated. The reactance of a capacitor is given by the formula:

$$X_C = 1 / (2 \pi f C)$$

Ohm's Law can be applied using the magnitude of a reactance in place of a resistance

$$V = I |X| \quad \text{or} \quad |X| = V / I \quad \text{or} \quad I = V / |X|$$

The current flowing through a capacitor *leads* the voltage across the capacitor by 90º. Conversely, the voltage across a capacitor *lags* the current flowing through the capacitor by 90º.

For capacitors connected in parallel,

$$C_{Total} = C_1 + C_2 + \ldots$$

While for capacitors in series,

$$1/C_{Total} = 1/C_1 + 1/C_2 + \ldots$$

There are many different types of capacitors suited to different purposes. Electrolytic capacitors are usually polarised, and one terminal must always remain positive with respect to the other. *Variable capacitors* were once used as the tuning control in many radios.

Revision Questions

1 The phase shift between voltage and current in a capacitor is:

a. 90°
b. 45°
c. 360°
d. 0°

2 Three capacitors of 1 µF each are connected in parallel. The equivalent capacitance is:

a. 330 nF
b. 3 µF
c. 300 nF
d. 33,33 µF

3 A capacitor of 250 pF is required to resonate a tuned circuit. A 100 pF capacitor is connected in parallel to a variable capacitor. What value must the variable capacitor be set to to achieve resonance?

a. 150 pF
b. 300 pF.
c. 350 pF
d. 400 pF

4 A value of 1000 pF is equal to:

a. 10 nF
b. 1 nF
c. 0,1 nF
d. 100 nF

5 The energy in a charged capacitor is stored in the:

a. Voltage across the terminals.
b. Current applied to the capacitor.
c. The electric field between the plates.
d. Form of magnetism.

6 The unit of capacitance is called:

a. farad
b. permeability
c. conductance
d. impedance

7 What is the total capacitance of two similar capacitors connected in parallel?

a. The same as either capacitor.
b. Half the capacitance of either capacitor.
c. Twice the capacitance of either capacitor.
d. The capacitance cannot be determined without knowing the exact values of the capacitors.

8 What do the units µF and pF specify?

a. Inductance.
b. Capacitance.
c. Resistance.
d. Current.

9 As the plate area of a capacitor increases, its capacitance:

a. Decreases.
b. Increases.
c. Stays the same.
d. Becomes voltage dependent.

10 Which of the factors below would *not* influence the capacitance value of a capacitor?

a. Area of the plates.
b. Distance between the plates.
c. Voltage rating.
d. Dielectric constant of the material between the plates.

11 The magnitude of the reactance of a capacitor:

a. Remains constant with changing frequency.
b. Increases with increasing frequency.
c. Decreases with increasing frequency.
d. Increases with decreasing frequency.

12 The capacitive reactance of a 16 µF, 40 V electrolytic capacitor to a signal of 100 Hz is:

a. 1 kΩ
b. 10 kΩ
c. 10 Ω
d. 100 Ω

13 If the frequency of AC applied to a capacitor is doubled, the capacitor's capacitive reactance will be:

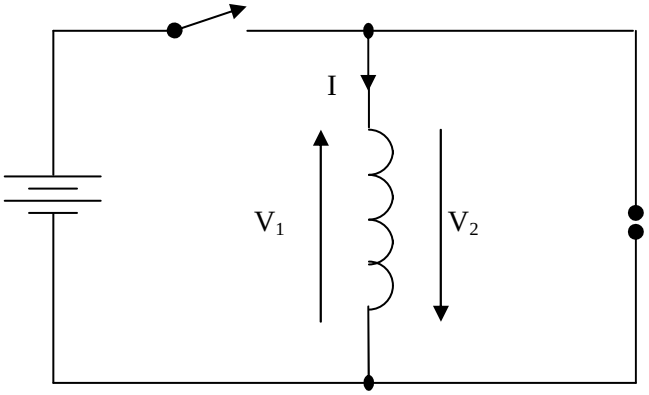
a. doubled.
b. four times original value.
c. one quarter the original value.
d. halved.

Chapter 10: Inductance and the Inductor

10.1 Inductors

A typical inductor consists of a coil of wire, which may be wound around a former or may be self-supporting. When a current flows through the wire, it generates a magnetic field, just like an electromagnet would. Whenever the current flowing through the inductor changes, the corresponding changes to the magnetic field induce a voltage into the inductor that opposes the change in the flow of current. This phenomenon is known as "self inductance" since the voltage is induced in the same coil that generates the magnetic field.

For example, consider the following circuit:



Circuit with inductor and spark gap

In this circuit, a battery is connected via a switch to an inductor. The inductor is represented by the component in the middle of the diagram that looks like a coil of wire. A spark gap is connected in parallel with the inductor, represented by the two dots on the right hand side of the diagram.

When the switch is closed, there is initially no current flowing in the inductor. However, the voltage of the battery will cause a current to flow, as the resistance of the wire making up the inductor is typically very low. This current causes the inductor to generate a magnetic field, and the growing magnetic field induces a voltage V_1 into the inductor that opposes the attempt to increase the current through the inductor. As a result, the current I flowing through the inductor will grow gradually, rather than reaching its full value as soon as the switch is closed.

When the switch is opened, the magnetic field starts to collapse, which induces a voltage V_2 across the inductor. V_2 acts to oppose the reduction in I that was initiated by opening the switch. Because there is no low-resistance path around the circuit with the switch opened, the only way it can allow current to flow is to generate a voltage that is high enough to cause a spark to jump across the spark gap. This induced voltage across an inductor, which is also called the *back EMF*, may be many times the supply voltage. Remember that EMF is the *electromotive force*, which we remember as an early term used to describe voltage.

The ignition circuit in cars with old (non-electronic) ignition systems works exactly this way. The ignition coil is an inductor, and the points act as a switch that opens, cutting off the current supply to the ignition coil and causing it to generate a high back EMF across one of the spark plugs. Even though most automotive electrical systems have a 12 V battery, the coil can generate a voltage of several kV across the spark plug.

When the switch is closed and current flows setting up a magnetic field, energy is taken from the circuit and stored by the inductor in its magnetic field. When the switch is opened, the inductor returns that energy to the circuit as its magnetic field collapses. So like a capacitor, an inductor "borrows energy from" and "returns energy to" the circuit, but does not actually dissipate power.

10.2 Inductor Values

The value of an inductor indicates how much energy it can store in its magnetic field, and hence how effectively it can oppose attempts to change the current flowing through it. The value of an inductor is measured in henry, with the abbreviation H. Typical values are measured in microhenry (μH) or millihenry (mH).

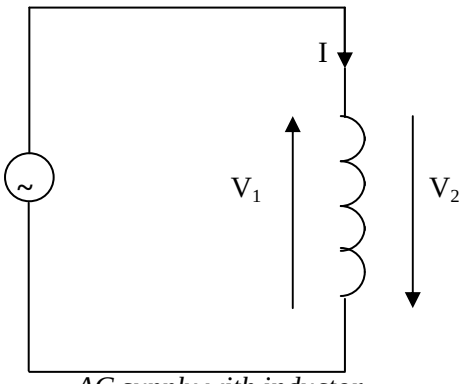
The value of an inductor depends on its physical characteristics:

- **The number of turns:** The more turns, the more inductance.
- **The coil diameter:** The larger the diameter, the more inductance.
- **The spacing between the turns of wire:** The closer the spacing, the more inductance. Therefore, a longer coil will have less inductance, all other factors being equal.
- **The permeability of the core:** The permeability of the core affects the strength of the magnetic field that will be caused by a current flowing through the inductor. Since ferrite has much higher permeability than air, a ferrite-cored inductor will have a greater inductance than an air-cored inductor with the same number of turns. Although ferrite cored inductors have higher inductance than air-cored inductors, they also have higher losses, especially at radio frequencies. Air core inductors may be wound with stiff wire, in which case they can be self supporting, or they may be wound on a plastic former.

Inductance is usually abbreviated "L", since "I" is already taken for current!

10.3 Inductors in AC Circuits

Consider the following circuit, which shows an AC voltage source connected to an inductor.



AC supply with inductor

The AC supply continually attempts to change the current flowing through the inductor. This current will change the magnetic field, which will in turn induce a voltage across the inductor that will oppose any change to the current flowing through the inductor.

For example, suppose the voltage source is attempting to increase the current flowing in the direction of *I*. The induced voltage will be in the direction V_1, opposing the increase in the current. However, when the voltage starts trying to reduce the current flowing in the direction *I*, the induced voltage will be in the direction of V_2, now opposing the attempt to reduce the current flowing in the direction of I, and so on.

The fact that the induced voltage always opposes the change in the flow of current does not mean that one cannot change the flow of current in an inductor. It just means that the current flowing through an inductor cannot change instantaneously; it will always take some time (depending on the value of the inductance and the voltage applied) to reach the final value.

10.4 Inductive Reactance

Because in AC circuits the current is always changing, and inductors oppose any attempt to change the current flowing through them, inductors oppose the flow of current in an AC circuit. However, this opposition is not resistance, since the inductor does not dissipate any power – it merely "borrows energy from" and "returns energy to" the circuit, just like a capacitor. As with capacitors, the opposition to the flow of current exhibited by an inductor in an AC circuit is *reactance*.

Consider the effect of frequency. The higher the frequency, the greater the rate at which the flow of current is changing. Since the inductor is effectively acting to oppose changes in the flow of current, it will exhibit higher reactance at high frequencies (where the current is changing fast) than at low frequencies (where current is changing slowly). This trend is evident in the formula giving the reactance of an inductor:

$$X_L = 2 \pi f L$$

where X_L is the reactance of the inductor in Ω, π the mathematical constant pi (approximately 3,14), f the frequency in Hz and L the inductance in H. It is proportional to the frequency: if the frequency is doubled, the reactance is doubled, and if the frequency is halved the reactance is halved.

10.5 Ohm's Law and Reactance

Once you have determined the reactance of an inductor, you can apply Ohm's Law to calculate the current or voltage in a circuit by replacing resistance with the magnitude of the reactance, |X|. For example, suppose a 1 V signal at a frequency of 1 MHz (10^6 Hz) is applied across an inductance of 10 μH (10^{-5} H). The reactance of the inductor *at this frequency* can be found as follows:

$$\begin{aligned} X_L &= 2 \pi f L \\ &= 2 \times 3{,}14 \times 10^6 \text{ Hz} \times 10^{-5} \text{ H} \\ &= 62{,}8 \ \Omega \end{aligned}$$

The current flowing through the inductor can be calculated using Ohm's Law, with resistance replaced by the magnitude of the reactance:

$$\begin{aligned} I &= V / |X| \\ &= 1 \text{ V} / 62{,}8 \ \Omega \\ &= 0{,}016 \text{ A} \\ &= 16 \text{ mA} \end{aligned}$$

Note that although reactance is measured in Ω, it is not the same as resistance, so resistances and reactances cannot be added together.

10.6 Phase Relationship between Voltage and Current

The voltage across an inductor always *leads* the current flowing through the inductor by 90º. Conversely, the current flowing through an inductor *lags* the voltage across the inductor by 90º. The 90º phase difference between the voltage and current means that no power is dissipated by a perfect inductor. Energy that is taken from the circuit and stored in

the magnetic field during one part of the cycle is returned to the circuit during another part of the cycle.

Real inductors are made of electrical wire that has some resistance. Although the resistance is usually small, some power is dissipated due to the resistance of the wire.

A useful acronym to remember the phase relationship of the voltages and currents in inductors and capacitors is “CIVIL”. The first three letters “CIV” mean “in a capacitor (C), current (I) leads voltage (V). The last three letters mean “voltage (V) leads current (I) in an inductor (L)”.

10.7 Inductors in Series and Parallel

Inductors in series and parallel behave similarly to resistors in series and parallel. For inductors in series,

$$L_{Total} = L_1 + L_2 + \ldots$$

while for inductors in parallel,

$$1/L_{Total} = 1/L_1 + 1/L_2 + \ldots$$

For example, if a 4,7 μH inductor is connected in parallel with a 3,3 μH inductor, the equivalent inductance could be found as follows:

$$\begin{aligned} 1/L_{Total} &= 1/L_1 + 1/L_2 + \ldots \\ &= 1/(4{,}7 \times 10^{-6}\ H) + 1/(3{,}3 \times 10^{-6}\ H) \\ &= 212\,766\ /H + 303\,030\ /H \\ &= 515\,796\ /H \end{aligned}$$

so $$\begin{aligned} L_{Total} &= 1 / 515\,796\ /H \\ &= 1{,}9\ \mu H \end{aligned}$$

Summary

Inductors store energy in their magnetic fields. As the current through an inductor changes, the changing magnetic field induces a voltage across the inductor that acts to oppose the change to the current flowing through the inductor. This phenomenon is called “self inductance”.

In AC circuits, inductors exhibit a reactance proportional to frequency. The formula for the reactance of an inductor is:

$$X_L = 2\,\pi\,f\,L$$

Ohm’s Law can be applied using the magnitude of a reactance in place of a resistance

$$V = I\,|X| \quad \text{or} \quad |X| = V / I \quad \text{or} \quad I = V / |X|$$

The voltage across an inductor *leads* the current flowing through the inductor by 90º. The phase relationships between voltage and current in capacitors and inductors can be remembered using the acronym “CIVIL”.

The equivalent inductance of two or more inductors in series is given by:

$$L_{Total} = L_1 + L_2 + \ldots$$

The equivalent inductance of two or more inductors in parallel is given by:

$$1/L_{Total} = 1/L_1 + 1/L_2 + \ldots$$

Revision Questions

1 The characteristic back-EMF which a collapsing magnetic field causes in a coil is called:

a. mutual inductance.
b. self inductance.
c. magnetic flux.
d. the solenoid effect.

2 What is the unit of inductance?

a. henry
b. coulomb
c. farad
d. ohm

3 A small air-core coil has an inductance of 5 µH. What do you have to do if you want a 5 mH coil with the same physical dimensions?

a. The coil must be wound on a non-conducting tube.
b. The coil must be wound on a ferrite core.
c. Both ends of the coil must be brought around to form the shape of a doughnut, or toroid.
d. The coil must be made of a heavier-gauge wire.

4 For radio frequency power applications, with which type of inductor would you get the least amount of loss?

a. Magnetic wire.
b. Iron core.
c. Air-core.
d. Slug-tuned.

5 In an inductive circuit, the alternating current produced in relation to the applied EMF is:

a. Lagging by 90°.
b. 180° out of phase.
c. Leading by 90°.
d. In phase.

6 The phase shift between voltage and current in an inductor is:

a. 90°
b. 45°
c. 360°
d. In phase.

7 The reactance of an inductor:

a. Remains constant with changing frequency.
b. Increases with increasing frequency.
c. Decreases with increasing frequency.
d. Increases with decreasing frequency.

Chapter 11: Tuned Circuits

Inductors and capacitors can be combined in series and parallel to form circuits that have the ability to accept or reject signals of particular frequencies. These circuits, which are called *tuned circuits*, are of great importance in radio.

As you will see later, modern radios use digital techniques to achieve tuning, but until late in the twentieth century, tuned circuits were used universally for selecting frequencies to receive.

11.1 Reactances in Series

Both capacitors and inductors exhibit *reactance* in AC circuits. The reactance depends on frequency according to the formulae:

$$X_C = 1 / (2 \pi f C)$$

and

$$X_L = 2 \pi f L$$

When reactances are connected in series – for example, two capacitors or a capacitor and an inductor – then the reactances can be added to give the equivalent reactance of the two reactances in series. However, remember that the phase lag between voltage and current is opposite for capacitors and inductors. To make provision for this difference, we add inductive reactances and subtract capacitive reactances to determine the total reactance.

$$X_L = X_{L1} + X_{L2} + \ldots$$

and

$$X_C = X_{C1} + X_{C2} + \ldots$$

If capacitive and inductive reactance are both present:

$$X_{Total} = X_L - X_C$$

For example, suppose we connect two 100 pF (10^{-10} F) capacitors in series. At a frequency of 10 MHz (10^7 Hz), the reactance of each of the capacitors individually is:

$$\begin{aligned} X_C &= 1 / (2 \pi f C) \\ &= 1 / (2 \times 3{,}14 \times 10^7 \times 10^{-10}) \, \Omega \\ &= 1 / 0{,}006\,28 \, \Omega \\ &= 159 \, \Omega \end{aligned}$$

So the equivalent reactance of the two reactances in series is:

$$\begin{aligned} X_{Total} &= X_1 + X_2 \\ &= -X_{C1} - X_{C2} \\ &= -159 \, \Omega - 159 \, \Omega \\ &= -318 \, \Omega \end{aligned}$$

Remember that capacitive reactance must be subtracted, to compensate for its opposite lag to that of an inductor.

Of course there is another way to find this result. Since we have two capacitors of the same value (100 pF) in series, the equivalent capacitance must be half the capacitance of the individual capacitors, or 50 pF (5×10^{-11} F). We can calculate the reactance of this equivalent 50 pF capacitance at 10 MHz (10^7 Hz) as follows:

$X_C = 1 / (2 \pi f C)$
$= 1 / (2 \times 3{,}14 \times 10^7 \times 5 \times 10^{-11}) \Omega$
$= 1 / 0{,}003\ 14\ \Omega$
$= 318\ \Omega$

or

$X_{Total} = -318\ \Omega$

As expected, we reach the same answer.

11.2 Reactances in Parallel

Similarly, the formula for the equivalent reactance of two reactances in parallel is:

$$1/X_{Total} = 1/X_1 + 1/X_2 + \ldots$$

For example, if we take our two 100 pF (10^{-10} F) capacitors, which each has a capacitive reactance of 159 Ω at 10 MHz, and connect them in parallel, the equivalent reactance is found as follows:

$1/X_{Total} = 1/X_1 + 1/X_2 + \ldots$
$= 1/-159\ \Omega + 1/-159\ \Omega$
$= -0{,}0126\ /\Omega$

so

$X_{Total} = 1/(-0{,}0126\ /\Omega)$
$= -79{,}5\ \Omega$

Once again this makes sense since the two 100 pF capacitors connected in parallel are equivalent to a single 200 pF (or 2 x 10^{-10} F) capacitor, with a reactance at 10 MHz of:

$X_C = 1 / (2 \pi f C)$
$= 1 / (2 \times 3{,}14 \times 10^7 \times 2 \times 10^{-10}) \Omega$
$= 1 / 0{,}0126\ \Omega$
$= 79{,}5\ \Omega$

11.3 The Series Tuned Circuit

Of course you might well ask, why bother to learn the formulas for reactances in series and parallel if we can calculate the same results using the formulas for capacitors and inductors in series and parallel that we already know? Good question; the answer can be found in the following circuit, which shows an inductor and a capacitor connected in series.

6,5 µH 39 pF

LC series tuned circuit

Suppose we want to calculate the equivalent total reactance of these two components at 10 MHz (10^7 Hz). We can't use the formula for inductors in series or the formula for capacitors in series, since the circuit contains one of each. So instead we must calculate the individual reactances of each component at a frequency of 10 MHz, and then use the formula for reactances in series.

The reactance of the inductor is found as follows:

$$X_L = 2 \pi f L$$
$$= 2 \times 3{,}14 \times 10^7 \times 6{,}5 \times 10^{-6}\ \Omega$$
$$= 408\ \Omega$$

The reactance of the capacitor is given by:

$$X_C = 1 / (2 \pi f C)$$
$$= 1 / (2 \times 3{,}14 \times 10^7 \times 39 \times 10^{-12})\ \Omega$$
$$= 1 / 0{,}006\ 908\ \Omega$$
$$= 408\ \Omega$$

So the combined reactance of the inductor and capacitor in series at 10 MHz is

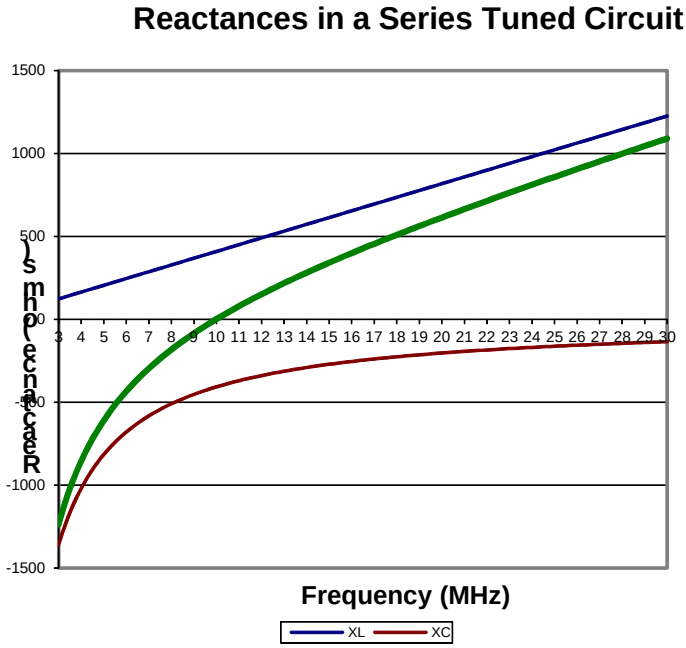
$$X_{Total} = X_L - X_C$$
$$= 408\ \Omega - 408\ \Omega$$
$$= 0\ \Omega$$

That's right—zero! The capacitor has reactance, and the inductor has reactance, but at this frequency (10 MHz) the positive reactance of the inductor exactly cancels out the negative reactance of the capacitor, leaving no reactance at all! The frequency at which the positive and negative reactances cancel out is known as the *resonant frequency* of the circuit. The circuit itself is called a *series resonant* circuit or a *series tuned* circuit.

Since the reactance of the inductor *increases* with frequency, while the reactance of the capacitor *decreases* with frequency, this canceling out will only happen at one specific frequency. At any other frequency, the circuit will exhibit either inductive (positive) or capacitive (negative) reactance.

Real tuned circuits also contain a little resistance, due to the wire from which the inductor is wound and the leakage current of the capacitor. However, good quality components will result in very low resistance.

The graph below shows the inductive reactance X_L (which is always positive), capacitive reactance X_C (always negative) and the combined reactance of the series circuit X_S. As you can see, the combined reactance is negative (capacitive) below the resonant frequency of 10 MHz, and positive (inductive) above the resonant frequency.



Reactance in a series tuned circuit

The series tuned circuit is very useful in radio electronics as the low reactance near the resonant frequency means that current can easily flow in the circuit near this frequency; while the high reactance at other frequencies will oppose the flow of current at frequencies other than the resonant frequency. In this way, a series tuned circuit can be used to accept signals with frequencies near the resonant frequency, while rejecting other signals.

11.4 Impedance

We have now learned about resistance, capacitive reactance and inductive reactance, all measured in Ω and all opposing current flow. We also learned that there is a 90° phase shift between current and voltage with a pure reactance (either capacitive or inductive). Now let us try to relate the three quantities.

In real life, a component has an *impedance*, measured in Ω, which describes the ratio between voltage and current. The impedance can consist of resistance and reactance. In DC circuits, there is no reactance and the impedance is simply equal to resistance. In AC circuits, there may be some reactance and the impedance is a combination of resistance and reactance.

In the previous section, we learned that inductive reactance is always positive, while capacitive reactance is always negative. These polarities describe the positive and negative phase shift associated with the particular reactance.

In a series circuit containing inductors and capacitors, you can simply add all the inductive reactances and all the capacitive reactances to arrive at the total reactance:

$$X_{Total} = X_{L1} + X_{L2} + X_{L3} + \ldots - X_{C1} - X_{C2} - X_{C3} \ldots$$

You can also add the resistances:

$$R_{Total} = R_1 + R_2 + R_3 + \ldots$$

The ratio between the two determines the phase shift. If there is lots of resistance and little reactance, there will be little phase shift. If there is lots of reactance and little resistance, the phase shift will be close to 90°, with the sense being dictated by the sign of the reactance (positive for inductive and negative for capacitive reactance).

For maths wizards only!
If you are comfortable with complex numbers, you can think of resistance as the real component of impedance, and reactance as the imaginary part. The total impedance is then

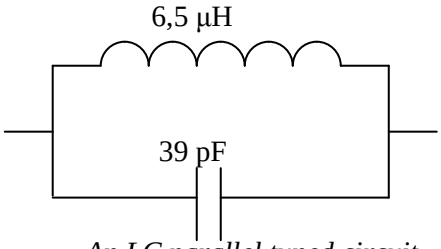
$$Z = R + jX = V / I$$

Where $j = (-1)^{½}$, which is orthogonal to 1 and which you may know from classical mathematics as *i*. The impedance is therefore the complex combination of *R* and *X*, and X_L and X_C are complex conjugates.

Obviously, the total phase shift between *V* and *I* is then dictated by the phase shift in *Z* (with $\Phi = tan^{-1} (X/R)$). All the forms of Ohm's law can then be comfortably used with *Z* instead of *R* and with voltage and current being represented as complex numbers or vectors.

11.5 The Parallel Tuned Circuit

Having seen the strange and interesting behaviour we get when we connect an inductor and capacitor in series naturally raises the question of what would happen if we were to connect them in parallel. To save us unnecessary calculations, we choose the same values of L = 6,5 µH and C = 39 pF.



An LC parallel tuned circuit

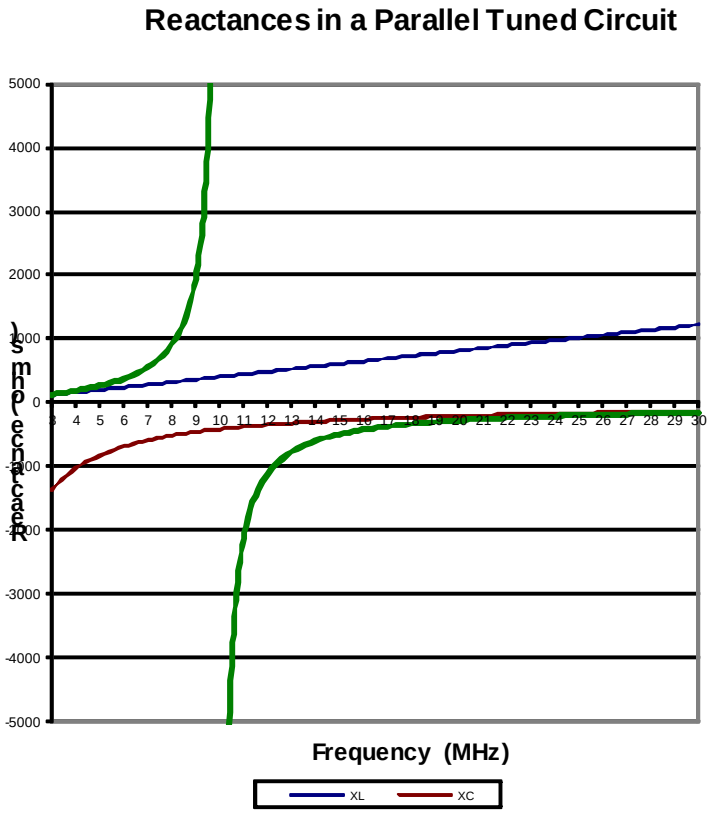
Once again, we will calculate the combined reactance at 10 MHz – since this was the resonant frequency for the series tuned circuit, perhaps it will also show some interesting behaviour in this *parallel tuned circuit*.

From the formula for reactances in parallel, we know that

$$1/X_{Total} = 1/X_L + 1/X_C$$
$$= 1/408\ \Omega + 1/-408\ \Omega$$
$$= 0{,}002\ 45\ /\Omega - 0{,}002\ 45\ /\Omega$$
$$= 0\ /\Omega$$

so $X_{EQUIV} = 1 / 0/\Omega$
$= ????$

What has happened here? Once again the positive inductive reactance has cancelled out the negative capacitive inductance, but this time it has left the zero in the denominator (bottom) of a fraction, which means that the result is undefined. However, if we plot a graph showing the reactances for a range of frequencies, we will understand what is happening better.



Reactance in a parallel tuned circuit

Once again the inductive reactance is always positive, while the capacitive reactance is always negative. This time, however, the combined reactance of the tuned circuit starts slightly positive (inductive) and rapidly gets more and more positive as the resonant frequency is approached. However, at the resonant frequency it instantaneously transitions from being a very high positive (inductive) reactance to being very high negative (capacitive) reactance. No wonder the exact value at resonance is undefined.

As a result, a parallel tuned circuit has a high reactance near resonance while its reactance is small away from the resonant frequency. This means that a parallel tuned circuit can be used to block signals near its resonant frequency, while allowing signals of other frequencies to pass relatively easily.

11.6 Circulating Current in a Parallel Tuned Circuit

A parallel tuned circuit has two components that are capable of storing energy. The inductor stores energy in its magnetic field; and the capacitor stores energy in the electric field between its plates. At resonance, energy is constantly being transferred from the capacitor to the inductor and back again.

As the capacitor charges up, a voltage develops between its plates. This voltage causes a current to flow through the inductor, which generates a magnetic field. As the capacitor discharges the voltage across its plates drops, which tends to reduce the current flowing through the inductor. However, an inductor will resist any attempt to change the current flowing through it. The magnetic field of the inductor collapses, inducing a potential difference into the inductor that acts to keep the current flowing in the same direction as it was before. This current flow now charges the capacitor up again, but with the opposite polarity to before. As the capacitor charges a voltage develops across its plates. This voltage causes current to flow through the inductor in the reverse direction, which generates a magnetic field, and so on.

So the parallel tuned circuit acts somewhat like a pendulum, continually transferring energy between two different forms. In the pendulum, these forms are the potential energy when the pendulum is stationary at the top of its arc, and the kinetic energy when the pendulum is moving at maximum speed at the bottom of its arc.

Remember that the parallel tuned circuit has lots of reactance at the resonant frequency. It therefore does not allow a lot of current to flow from the surrounding circuit. However, the *circulating current* that flows in a parallel tuned circuit – that is, the current flowing around the circuit containing the capacitor and the inductor – can be much larger. In practical circuits, it is not uncommon to have a circulating current that is 100 times the input current.

11.7 Calculating the Resonant Frequency

We have seen that in both a *series tuned circuit* and a *parallel tuned circuit*, something interesting happens at the *resonant frequency* which is where the reactance of the capacitor and inductor have the same magnitude (value) but one is positive and the other is negative. The reactances counteract one another. We can derive a formula for the resonant frequency as follows:

At resonance, the magnitude of the capacitive and inductive reactances are equal, so

$$X_L = X_C$$
$$2 \pi f L = 1 / (2 \pi f C)$$

so

$$f^2 = 1 / (4 \pi^2 L C)$$

and

$$f = 1 / (2 \pi \sqrt{LC})$$

You do not need to know the derivation, but you should be able to apply the result. For example, let us calculate the resonant frequency of a series or parallel circuit consisting of a 6,5 μH inductor and a 39 pF capacitor:

$$f = 1 / (2 \pi \sqrt{LC})$$
$$= 1 / (2 \times 3{,}14 \times \sqrt{6{,}5 \times 10^{-6} \times 39 \times 10^{-12}}) \text{ Hz}$$
$$= 1 / (6{,}28 \times \sqrt{253{,}5 \times 10^{-18}}) \text{ Hz}$$
$$= 1 / (6{,}28 \times 1{,}59 \times 10^{-8}) \text{ Hz}$$
$$= 1 / 10^{-7} \text{ Hz}$$
$$= 10^{7} \text{ Hz}$$
$$= 10 \text{ MHz}$$

This answer agrees with the resonant frequency in the series and parallel resonant circuits above.

11.8 Circuit Losses and the Quality Factor

The discussion so far has ignored circuit losses. For example, all practical inductors have some resistance as well as their inductance, and capacitors also have some losses although these are typically negligible compared to the losses caused by the resistance of the inductor.

The effect of these losses is that in a practical series tuned circuit, although at resonance the *reactance* would be zero, there would still be some small *resistance*. In a parallel tuned circuit, the effect of circuit losses is to limit the reactance at resonance to a high but finite value, rather than being completely undefined (or "infinite") as predicted by the maths.

The extent of circuit losses is expressed by a number called the "Quality Factor", or "Q Factor" or just the "Q" of the tuned circuit. A high Q means low circuit losses, while a low Q means high circuit losses. The Q is defined as the reactance of either the inductor or the capacitor at resonance (remember they are equal?) divided by the circuit resistance. So

$$Q = X_L / R$$
$$= X_C / R$$

The Q of practical tuned circuits is typically between 50 and 200.

The Q is related to two other properties of the tuned circuit:

1. The ratio of circulating current in a parallel tuned circuit to the current drawn by the tuned circuit is the same as the Q. So in a parallel tuned circuit with a Q of 100, the circulating current will be 100 times greater than the current drawn from the rest of the circuit.

2. The selectivity of the circuit – that is, its ability to allow desired signals through while blocking undesired signals. The greater the Q of the tuned the circuit, the greater its selectivity.

Summary

The series tuned circuit has a low reactance near its resonant frequency and a high reactance at other frequencies. Series tuned circuits are often used to allow signals near the resonant frequency to pass, while blocking signals at other frequencies.

Impedance is in AC circuits what resistance is in DC circuits. Impedance *Z* is composed of resistance *R* and reactance *X*. The ratio between *R* and *X* determines the phase shift caused by an impedance.

The parallel tuned circuit has a high reactance near its resonant frequency and a low reactance at other frequencies. Parallel tuned circuits are often used to block signals near the resonant frequency, while allowing signals at other frequencies to pass.

The resonant frequency of a series or parallel tuned circuit may be calculated as

$$f = 1 / (2 \pi \sqrt{LC})$$

The Quality Factor ("Q") is defined as the reactance of either the inductor or the capacitor at resonance divided by the circuit resistance. A tuned circuit with a high Q is more selective than a tuned circuit with a low Q.

The circulating current in a parallel tuned circuit may be many times the current drawn by the tuned circuit. The ratio between the circulating current and the current flowing into the tuned circuit is the same as the Q.

Revision Questions

1 At one particular frequency, resonance of a capacitor and inductor takes place. At this frequency:

a. Inductive reactance is nil.
b. Capacitive reactance is nil.
c. The impedance is nil.
d. The capacitive and inductive reactances are equal.

2 The parallel tuned circuit impedance at resonance is:

a. Low.
b. High.
c. Infinitely high.
d. Equal to 10.

3 The series tuned circuit impedance at resonance is:

a. Low.
b. High.
c. Infinitely high.
d. Equal to 10.

4 The Q of a parallel resonant circuit determines the:

a. Losses of the circuit.
b. Value of the capacitance required for resonance.
c. The inductor value required for resonance.
d. Value of increased current through the coil and capacitor at resonance.

5 The selectivity of a resonant circuit is greater if the Q factor:

a. Is low.
b. Decreases to 1.
c. Is high.
d. Remains low.

6 The resonant frequency of a tuned circuit consisting of a 10 nF capacitor in parallel with a 10 µH inductor is approximately:

a. 500 kHz
b. 5 MHz
c. 50 MHz
d. 500 MHz

7 You have a 100 µH inductor and wish to create a tuned circuit with a resonant frequency of 3,5 MHz. What value of capacitor would you require?

a. 2,1 pF
b. 12 pF
c. 21 pF
d. 120 pF

8. You have a 10 pF capacitor and wish to create a tuned circuit with a resonant frequency of 10 MHz. What value of inductor do you require?

a. 2,5 µH
b. 10 µH
c. 25 µH
d. 100 µH

Chapter 12: Decibel Notation

12.1 The Decibel

In amateur radio we often deal with ratios of powers. For example, the *gain* of an amplifier is the ratio of its output power to its input power. These ratios can be very large or very small. For example, the gain of a typical amateur radio receiver—the ratio between the output power into the speaker or headphones to the input power from the antenna—is in the region of 100 000 000 000 000. That's an amplification of a hundred trillion times! While we could use scientific notation to represent these large numbers (the one above is 10^{14}), another way of expressing the ratio of two powers is commonly used: the "decibel".

The unit "bel" was first used by telephone engineers at Bell Laboratories (now part of Alcatel-Lucent) and was named after Alexander Graham Bell (1847 - 1922), the inventor of the telephone and founder of Bell Laboratories. The "decibel" is simply one tenth of a bel, which turned out to be a more useful size. One decibel represents roughly the minimum discernable change in the loudness of an audio signal. The abbreviation for the decibel is "dB", which is also often used in general conversation such as "your signal is S9 plus 20 dB".

The decibel has great practical value, as it expresses ratios in a form that is very similar to how our ears perceive those ratios.

A ratio of two powers can be expressed in decibels as follows:

$$R_{dB} = 10 \log_{10} (R_P)$$

where R_P is the ratio of two powers (e.g. $R_P = P_1/P_2$), "R_{dB}" is the same ratio expressed in decibels, and "$\log_{10}$" means the mathematical logarithm to the base 10. If you are not familiar with logarithms, don't panic. Once we have explored a couple of the properties of decibels we will see that there is a simple way to calculate many common values.

12.2 Adding Decibels

A fundamental property of decibels (like any other logarithm) is that when two ratios expressed in decibels are added, it is equivalent to multiplying the original ratios. For example, a ratio of 2 times is 3 dB and a ratio of 10 times is 10 dB. If we add the decibel representations we get 3 dB + 10 dB = 13 dB, which is equivalent to a ratio of 2 x 10 = 20 times. This bit of magic is possible because of the use of the logarithm function in the definition of the decibel.

Example

In a radio receiver the radio frequency (RF) amplifier has a gain of 6 dB; the intermediate frequency (IF) amplifier has a gain of 110 dB and the audio frequency (AF) amplifier has a gain of 20 dB. What is the total gain of the receiver?

If the gains of the amplifiers had been expressed as simple ratios (P_{Out}/P_{In}), we would have to *multiply* the ratios together to get the total gain. However, since the gains are expressed in decibels, we can *add* them to get the total gain. So in this case the total gain is 6 dB + 110 dB + 20 dB = 136 dB.

12.3 Representing Losses

The decibel can also be used to represent losses, i.e. situations where a signal gets smaller. If you calculate the decibel equivalent of a ratio that is less than 1, the formula gives a

negative number. For example we can calculate the decibel equivalent of a power ratio of 0,1 (or one-tenth) as follows:

$$\begin{aligned} R_{dB} &= 10\ log_{10}\ (R_P) \\ &= 10\ log_{10}\ (0,1) \\ &= 10 \times -1\ dB \\ &= -10\ dB \end{aligned}$$

So, for example, an attenuator that reduces a signal to one-tenth its original power could be described as having a *gain* of –10 dB. Note that the minus sign indicates that it is actually making the signal smaller even though it is expressed as a "gain". The same attenuator could also be described as having a *loss* of 10 dB. This time there is no minus sign because it is being described as a *loss*.

However, if you add decibels together (which as we have seen is equivalent to multiplying the original ratios), you should express all the ratios as either gains or losses before adding them together. You can't add a decibel representing a *gain* to one representing a *loss*.

Example

An attenuator with a loss of 6 dB is added before the RF amplifier in a receiver. Before adding the attenuator, the receiver had a gain of 136 dB. What is the total gain of the receiver with the attenuator?

Because we can't add the 6 dB *loss* of the attenuator to the 136 dB *gain* of the receiver, we first convert express the attenuator's *gain* as –6 dB. Then we can calculate the total gain of the receiver by adding the –6 dB gain of the attenuator to the 136 dB gain of the receiver to get the answer 130 dB.

Finally, a gain of exactly 1 (i.e. a signal that gets neither stronger nor weaker) can be represented as 0 dB. *Adding* 0 dB to a ratio represented in decibels will not change it; just as *multiplying* a ratio by 1 won't change it either.

12.4 Quick and Easy Decibel Conversions

Some commonly used ratios are easily converted to decibels. These are shown in the table below:

Ratio	Decibels	Ratio	Decibels
1 000 000	60 dB	0,000 001	-60 dB
100 000	50 dB	0,000 01	-50 dB
10 000	40 dB	0,0001	-40 dB
1000	30 dB	0,001	-30 dB
100	20 dB	0,01	-20 dB
10	10 dB	0,1	-10 dB
5	7 dB	0,2	-7 dB
4	6 dB	0,25	-6 dB
2	3 dB	0,5	-3 dB
1	0 dB	1	0 dB

You don't need to remember all the powers of ten (the numbers 10, 100, 1000 etc.). If a ratio consists of a 1 followed by any number of zeros, you can convert it to decibels by simply multiplying the number of zeros by ten. For example, 1 000 000 has 6 zeros so it is equivalent to 60 dB (the number of zeros times ten).

Using these values it is possible to easily calculate the decibel representation of many other common ratios. For example, what is the decibel equivalent of a ratio of 20:1? 20 is not in the table, but 2 and 10 are, and 20 = 2 x 10. However, we know that multiplying ratios is the same as adding their decibel equivalents, so the decibel equivalent of 20 must be the decibel equivalent of 2 plus the decibel equivalent of 10. So the answer is 3 dB + 10 dB = 13 dB, which is the decibel equivalent of 20.

Of course we could use the thinking in the opposite direction too. Suppose we want to calculate the ratio represented by 27 dB. Although 27 dB is not in the table, we know that 27 dB = 20 dB + 7 dB, and both the values *are* in the table. Since adding decibels is equivalent to multiplying ratios, the ratio represented by 27 dB is the ratio represented by 20 dB *multiplied by* the ratio represented by 7 dB. So the answer is 100 x 5 = 500, which is the ratio represented by 27 dB.

We could, of course, also have used the fact that 27 =30 – 3. A ratio of 27 dB is therefore equivalent to 1000 ÷ 2 = 500.

12.5 Expressing Voltage Ratios as Decibels

Throughout this module I have stressed that decibel notation is used to express the ratio of two *powers*. However, because there is a relationship between voltage and power, decibels are also sometimes used to express the ratio between two *voltages*. Of course, each of these voltages is associated with a certain amount of power, determined by the voltage and the resistance. The decibel ratio must obviously provide the same answer, regardless of whether we are using the voltages or the powers to calculate the ratio.

The relationship between voltage and power can be expressed as

$$P = V^2 / R$$

Because power is proportional to the voltage *squared*, if the voltage is doubled then the power will be multiplied by 4; if the voltage is increased by a factor of 10 then the power will be multiplied by 100. This ratio will hold true for any resistance as long as the same resistance is used to calculate the power before and after the voltage is increased.

As a square involves multiplying a number by itself (implying that the ratio in decibels is doubled), we can modify our original formula to express a voltage ratio in decibels:

$$R_{dB} = 10 \log_{10} R_V^2$$

$$= 2x\ 10 \log_{10} R_V$$

$$= 20 \log_{10} R_V$$

where R_V is the ratio of two voltages and R_{dB} is the same ratio expressed in decibels. Note that the constant “10” in the formula used for power ratios is replaced by “20” in the formula for voltage ratios. This is to take into account the V^2 factor in the formula for power. In other words, when we representing a voltage ratio in decibels, we are still representing a ratio between two powers. In this case, however, it is the notional power that would be dissipated by some (unknown) load if the voltages in question were applied across the load.

Note that expressing voltage ratios as decibels is a confusing and potentially misleading exercise. Wherever possible, deal with *power* ratios not voltage ratios.

For example, suppose the input voltage of an amplifier is 10 μV and the output voltage is 1 mV. The input and output resistances of the amplifier are both 50 Ω and we want to calculate the gain of the amplifier in decibels.

The input and output powers can be found from

$$P_{In} = V^2 / R$$
$$= (10 \times 10^{-6})^2 / 50\ W$$
$$= 2\ pW$$

$$P_{Out} = V^2 / R$$
$$= (10^{-3})^2 / 50\ W$$
$$= 20\ nW$$

Having calculated the powers, we can express them as a ratio and then convert the ratio to decibels:

$$P_{Out}/P_{In} = 20\ nW / 2\ pW$$
$$= 10\ 000$$
$$= 40\ dB$$

An alternative way to reach the same answer would be to take the voltage ratio

$$R_V = 1\ mV / 10\ \mu V$$
$$= 100$$

Then square this to find the power ratio

$$R_P = 100^2$$
$$= 10\ 000$$

And then express this ratio as 40 dB four zeros, multiplied by ten!) However, this alternative approach will only work if the input and output resistances are equal. The first method – calculating the actual input and output powers – will work whatever the input and output resistances, as long as you know what they are.

12.6 Expressing Power Levels in dBW and dBm

In the new Radio Regulations, the power levels that apply to amateur transmissions are not expressed in watts as before, but rather in dBW. The unit dBW means “decibels referenced to 1 W”. It is a way to express actual powers in decibel notation. Note that one cannot express an actual power – say 100 W – in decibels since decibels are used to express the *ratio* of two powers. However, if you make one of the two powers a standard reference level, by expressing the ratio of the other power to this standard reference level you can communicate an actual power level. One of the common reference levels is 1 W, and the resulting unit is given the abbreviation “dBW”. For example, the maximum power level specified on some frequency bands is 26 dBW. This means “26 dB higher than 1 W”. Since 26 dB is a ratio of 400, 26 dBW means 400 W.

A related unit is decibels over 1 mW, abbreviated “dBm”. For example, the sensitivity of most amateur receivers is around –130 dBm, meaning “130 dB less than 1 mW”. This power is an incredibly small 10^{-16} W, or 0,0001 pW!

Summary

The decibel is a logarithmic unit used to express the ratio of two powers. The ratio of two powers can be converted to decibels using the formula

$$R_{dB} = 10 \log_{10} (R_P)$$

Adding two ratios expressed in decibels is equivalent to multiplying the original ratios. However, both of the figures added must express either a gain or a loss; you cannot add a gain to a loss. To convert a gain to a loss or *vice versa*, simply put a minus sign before it. If a ratio consists of a 1 followed by any number of zeros, to convert it to decibels you simply multiply the number of zeros by ten.

A ratio of two voltages can be expressed in decibels using the formula

$$R_{dB} = 20 \log_{10} R_V$$

This formula reflects the fact that the voltage ratio can be converted to a power ratio by squaring it, and then expressing the resulting power ratio in decibels. This will only give the correct result if both voltages are applied across the same resistance.

Although absolute powers cannot be expressed in decibels, they can be expressed in dBW (decibels referenced to 1 W) or dBm (decibels referenced to 1 mW).

Revision Questions

1 An increase in power from 250 mW to 1,25 W is equal to a gain of:

a. 3 dB
b. 7 dB
c. 10 dB
d. 1 dB

2 A transmitter with a power output of 100 W is connected to an antenna with 11 dB gain by means of a coax cable with a loss of 1 dB. The ERP (effective radiated power) of the transmitter, coax and antenna combined is:

a. 11 W
b. 111 W
c. 1 kW
d. 2 kW

3 A 20 dB attenuator is placed in line with a 40 V RMS signal. Assuming the impedances all remain constant what will the reduced signal level be?

a. 2 V
b. 10 V
c. 20 V
d. 4 V

4 A power gain of 4 is equivalent to:

a. 3 dB
b. 6 dB
c. 10 dB
d. 16 dB

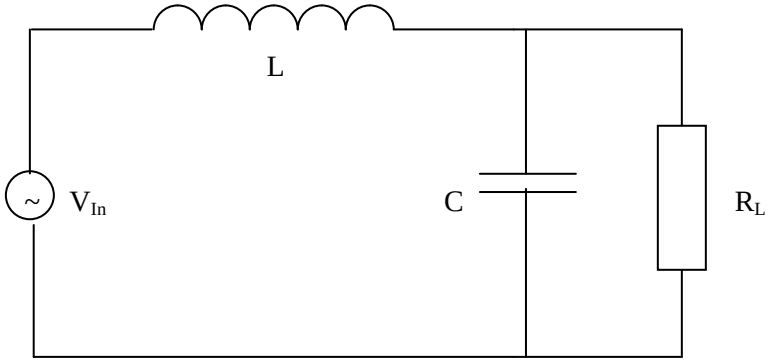
5 A signal with a power of 1 mW is applied to the input of an amplifier that has a gain of 13 dB. The power of the output signal will be:

a. 5 mW
b. 10 mW
c. 20 mW
d. 100 mW

Chapter 13: Filters

Filters are electrical circuits that allow signals of particular frequencies to pass, while blocking signals of other frequencies. They can be used, for example, to select the signal that a radio receiver is tuned to, while blocking all signals that it is not tuned to.

13.1 The Lowpass Filter



LC lowpass filter

An input voltage V_{In} is applied across a voltage divider consisting of an inductor L and a capacitor C in parallel with a resistive load, R_L.

Although we are not in a position to analyze this circuit quantitatively, we can get a good qualitative idea of what happens. We think about what would happen at different frequencies, including extreme cases.

At DC (i.e. the frequency is 0 Hz!), the inductor would behave like a short circuit and the capacitor like an open circuit. The full source voltage will therefore appear across the load.

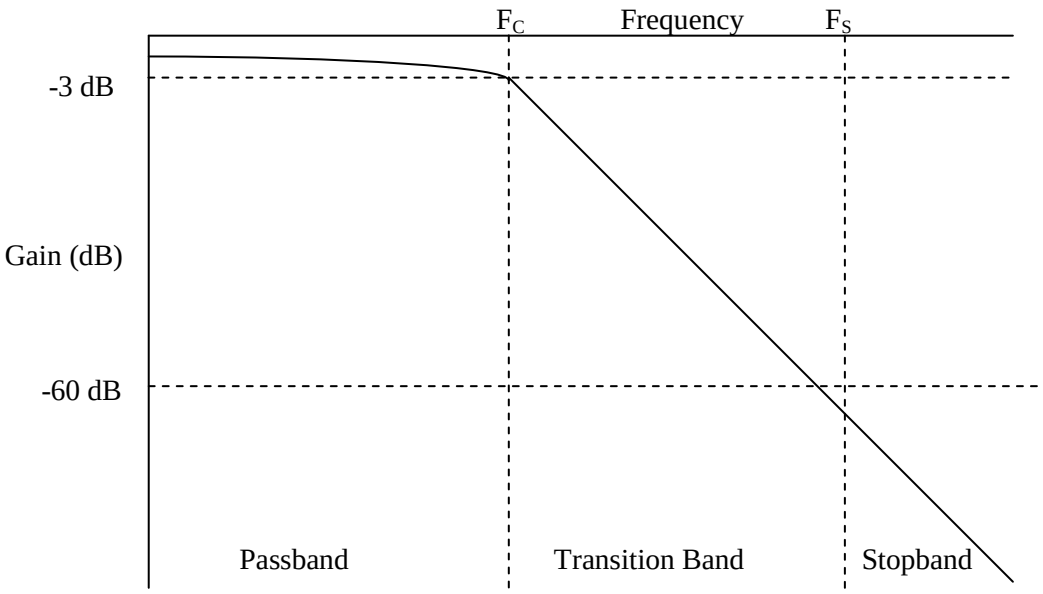
When the frequency of the input voltage is low, the inductor has low reactance while the capacitor has high (negative) reactance. There is little opposition to current flowing through L, but significant opposition to current flowing through C. As a result, most of the input voltage is applied across the load resistance R_L, and power is efficiently transferred to the load.

At a higher frequency, since the reactance of an inductor is proportional to the frequency, L will have high reactance. On the other hand, the reactance of a capacitor decreases with frequency, so C will have a low impedance. This means that the inductor provides significant opposition to the flow of current; and what current is able to flow is mostly diverted through the capacitor rather than flowing through the load. As a result, little power is transferred to the load.

At some very high frequency, and disregarding phase lag, the inductor will behave like an open circuit (very high reactance) and the capacitor like a short circuit (very low reactance), providing no power to the load.

This circuit is called a “lowpass filter” because it allows low frequency signals to pass (in other words, to be efficiently coupled to the load), while blocking high frequency signals.

A graph can be plotted showing the *frequency response* of the filter – that is, its gain at different frequencies.



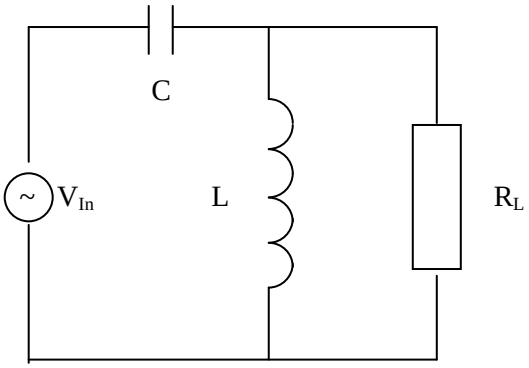
The Frequency Response of a Lowpass Filter

The *cutoff frequency* F_C is the frequency at which the attenuation of the filter is 3 dB (i.e. the gain is –3 dB). At this frequency, half the input power reaches the load. For a lowpass filter, signals with frequencies lower than the cut-off frequency have relatively little attenuation; these signals are in the *passband* of the filter.

Signals with frequencies higher than F_S are greatly attenuated – in this case by 60 dB or more. These signals are in the *stopband* of the filter. Signals with frequencies between F_C and F_S are somewhat attenuated. These frequencies are sometimes called the *transition band* of the filter since it is in transition between the passband and the stopband.

Most amateur radio transmitters have a lowpass filter after the final power amplifier to attenuate any *harmonics* of the output frequency. Harmonics are multiples of the output frequency caused by distortion in the amplifier, so for example a transmitter that is transmitting on a frequency of 3,5 MHz might have harmonics on 7 MHz, 10,5 MHz, 14 MHz, 17,5 MHz, 21 MHz and so on. It is very difficult to design a power amplifier that does not generate any harmonics, and in any case such an amplifier would probably be quite inefficient. However, it is easy to use a lowpass filter at the output to pass the desired frequencies and attenuate the harmonics to an acceptably low level.

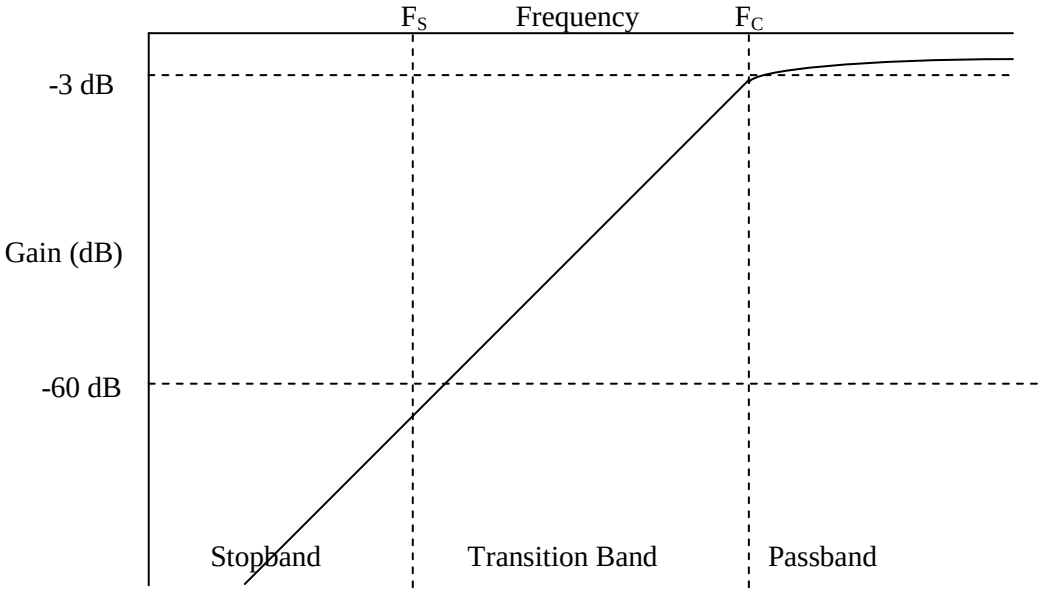
13.2 The Highpass Filter



Highpass LC filter

Once again the input voltage V_{In} is applied to a voltage divider, but this time the capacitor and inductor in the voltage divider have been swapped. At low frequencies, the capacitor has high reactance and so opposes the flow of current; while the inductor has low reactance so the current that does flow is diverted through the inductor rather than flowing through the capacitor.

At high frequencies, the capacitor has low reactance, so does little to oppose the flow of current. The inductor has high reactance, so most of the current flows through the load resistor R_L rather than through the inductor. This circuit is called a "highpass" filter because it allows high frequency signals to pass (in other words to be efficiently coupled to the load) while blocking low frequency signals, including DC. The frequency response of a highpass filter looks something like this:



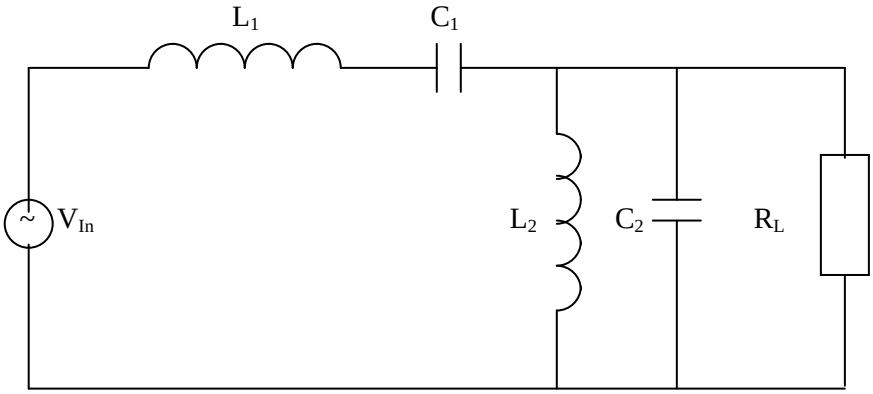
The Frequency Response of a Highpass Filter

Once again, the cutoff frequency is the frequency at which the attenuation of the filter is 3 dB (the *half-power* point), while I have chosen to measure the stopband from the point where the attenuation is 60 dB.

Highpass filters are often used in the input stages of receivers to reject the very strong radio signals found in the medium wave broadcast band between 500 kHz and 1,5 MHz so they do not overload the receiver, while allowing signals in the amateur bands starting at 1,8 MHz to pass.

13.3 The Bandpass Filter

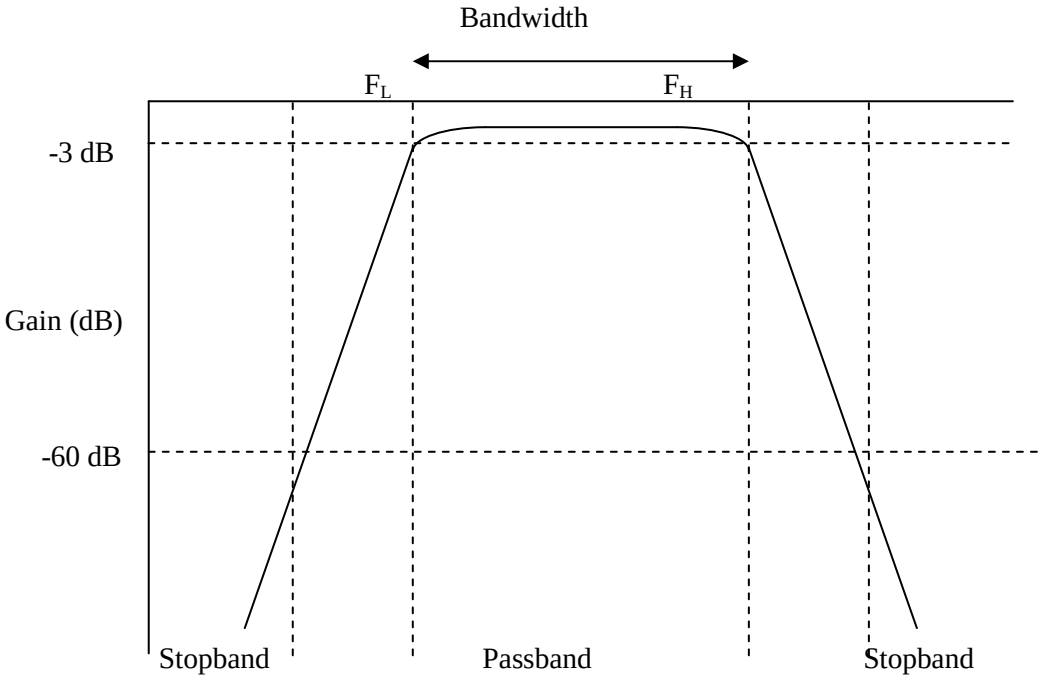
Bandpass filters pass signals in a certain frequency range known as the *passband* and reject signals with frequencies above or below the passband. They can be constructed using both series and parallel tuned circuits. For example, consider the circuit below:



LC Bandpass Filter

Once again we have a circuit resembling a voltage divider, although this time it is made up of two tuned circuits – a series tuned circuit consisting of L_1 and C_1 in series with the source, and a parallel tuned circuit consisting of L_2 and C_2 across the load. Assume that the two tuned circuits have the same resonant frequency. Near this frequency, the series tuned circuit has low reactance while the parallel tuned circuit has very high reactance, so almost the entire input voltage appears across the load. This frequency range is the passband of the filter.

However, at frequencies well above or below the resonant frequency, the series tuned circuit has a high impedance while the parallel tuned circuit has a low impedance, so very little of the input voltage appears across the load. This is the stopband of the filter.



The Frequency Response of a Bandpass Filter

The bandpass filter has two cutoff frequencies, a high cut-off labeled F_H and a low cut-off labeled F_L. Both cutoff frequencies are measured at the point where the output from the filter is 3 dB below the input to the filter (the *half-power* points). The *bandwidth* of the filter is the difference (in Hz) between the high cutoff frequency and the low cutoff frequency. For example, if the high cutoff frequency is 2700 Hz and the low cutoff frequency is 300 Hz, the bandwidth is 2700 Hz – 300 Hz = 2400 Hz. The *centre frequency*

of a bandpass filter is the frequency half way between the high cutoff frequency and the low cutoff frequency; in this case 1500 Hz.

Most amateur receivers use bandpass filters to allow signals from a particular amateur band to enter the receiver while rejecting signals from other amateur bands. Such filters are called *preselectors*.

Bandpass filters are generally specified with a *bandwidth*, a *slope factor* and an *ultimate rejection* figure. The bandwidth is normally between the -6 dB points, while the slope factor is defined as the ratio of the bandwidth and the spacing between the -60 dB cutoff points. A filter with a bandwidth of 500 Hz and a -60 dB cutoff spacing of 4 kHz has a slope factor of 8. The attenuation in the stopband might be 94 dB, which is the ultimate rejection of the filter.

A good quality filter has the desired bandwidth (as appropriate for the signals to be passed), a small slope factor and very high ultimate rejection.

13.4 Crystal Filters

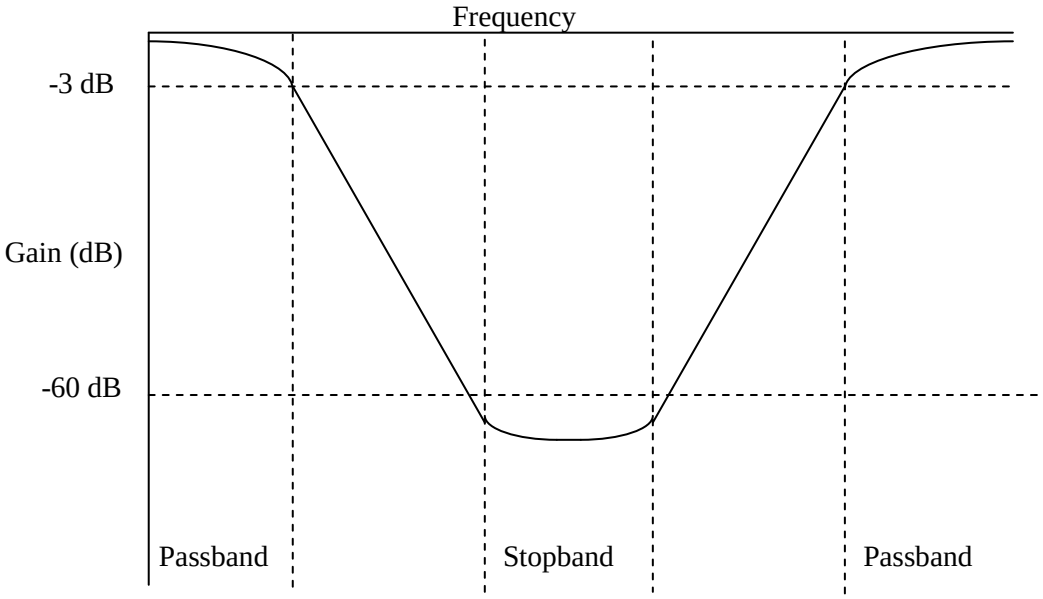
Bandpass filters can also be implemented using quartz crystals. Quartz has a piezoelectric property, which means that a voltage applied to the crystal causes a slight physical distortion of the crystal; and physical movements of the crystal will in turn cause a voltage to appear across it. Quartz crystals have very similar properties to tuned circuits and can be used to make highly selective bandpass filters. These “crystal filters” are responsible for the selectivity—that is, the ability to distinguish one signal from another—of many modern amateur receivers and transceivers.

Although crystal filters are very selective—that is, their bandwidth is very narrow in comparison with the centre frequency of the filter—they have the disadvantage that they only work at a single fixed frequency. A crystal filter cannot be tuned to different frequencies. When we look at the design of superhet receivers we will see how this limitation is overcome while allowing the receiver to take advantage of the exceptionally good selectivity of crystal filters.

Amateur receivers and transceivers often offer different bandwidth crystal filters for different purposes. Some of the common bandwidths are 2,4 kHz for normal phone (SSB) operation, 1,8 kHz for phone operation under difficult conditions (often used in contests) and between 250 Hz and 500 Hz for CW (Morse Code) operation. Most transceivers come with one or two basic filters (for example, just a 2,4 kHz filter) but additional filters can often be purchased, albeit at a price– they typically retail for US$ 100 to US$ 500 per filter.

13.5 The Bandstop Filter

A bandstop filter works in the opposite way to a bandpass filter. Frequencies in a certain range (the stopband) are attenuated, while frequencies either above or below those frequencies are passed. Amateur receivers and transceivers often provide a manually adjustable bandstop filter that can be used to attenuate undesired signals, for example a carrier generated by someone tuning up close to the frequency that you are listening to. These are known as “notch filters” because they allow you to “notch out” undesired signals.



The Frequency Response of a Bandstop Filter

13.6 More Sophisticated Filters

The filters shown here have all been simple filters, using one or two LC groups to achieve their objectives. In principle, several stages can be used in a row, providing even more of the same effect. Such filters are known as multi-pole filters. Using such filters, it is possible to get better bandwidth, better slope factors and better ultimate rejection. Complex filters may also introduce complications to the basic function of the filter, such as variation in the function of the filter in the intended passband, or unwanted phase shifts Where the filter attenuation varies in the desired passband, it is known as *passband ripple*.

Many standard filter configurations are in use, including Butterworth, Chebyshev, and Bessel. Each has its advantages and drawbacks in terms of insertion loss, passband ripple, impulse response, ultimate rejection, ringing and slope factor.

13.7 Practical RF Circuits

Perhaps this point in the syllabus is a good time to tackle the issue of non-ideal components.

Any practical component is not purely an inductor or resistor or capacitor.

Inductors have some resistance, which acts like a series resistor in a circuit. At high enough frequencies, they also have inter-winding capacitance, known as *stray capacitance*. At really high frequencies, the stray capacitance starts to override the inductance, and the inductor starts acting like a capacitor!

Resistors also have nasty characteristics. We have already mentioned the fact that wirewound resistors have lots of inductance, so we can expect them to start acting like an inductor once the frequency gets high enough. Even carbon-film resistors have stray inductance and capacitance.

Capacitors likewise have bad habits. Electrolytic capacitors, often used where large values are required in a compact design, such as in power supply filter circuits, cannot respond to quick changes in voltage. All capacitors have some stray inductance due to the leads.

Even interconnections, such as PC board tracks and point-to-point connecting wires, start exhibiting stray inductance and capacitance at a high-enough frequency. In fact, PCB tracks start acting like transmission lines. High-speed computer boards are designed by automated

software to take account of the many different effects exhibited by those tiny close-spaced tracks.

What is a “high frequency” in this discussion? The answer varies from component to component. Some components are made for UHF applications, and can be expected to behave pretty well even at 1 GHz or more. Others start degrading even at 1 MHz. The answer is in the ratio of resistance to reactance, or of capacitive to inductive reactance. When a resistor starts exhibiting phase shifts or reactance, it is no longer a resistor. Likewise, when a reactive component (L or C) starts exhibiting as much resistance as reactance, it is no longer doing its job.

When designing and building a circuit for radio frequencies, be mindful of these problems. Select components that have good characteristics at the design frequency. Make interconnections as short as possible. Ensure that connections are well soldered or crimped. And never underestimate the potential of RF to confound your attempts to find a problem!

Summary

Lowpass filters allow signals with frequencies below the cut-off frequency to pass with little attenuation, while significantly attenuating signals with frequencies well above the cut-off frequency. Highpass filters allow signals with frequencies above the cut-off frequency to pass with little attenuation, while significantly attenuating signals with frequencies well below the cut-off frequency. In both cases, the cut-off frequency is measured from the point where the signal is attenuated by 3 dB; also known as the “half power” point.

Bandpass filters allow signals with frequencies between the low and high cut-off frequencies to pass, while attenuating signals with frequencies significantly higher or lower than the passband. The bandwidth of a bandpass filter is the difference between the high cut-off and low cut-off frequencies. Crystal filters are highly selective bandpass filters. Bandstop filters attenuate signals with frequencies in a particular range, while allowing signals outside that frequency range to pass. Very narrow bandstop filters can be used as notch filters.

Good bandpass filters have a good slope factor and good ultimate rejection.

Filters can be ganged to achieve an enhanced effect. Multi-pole filters can achieve various combinations of insertion loss, ringing, passband ripple, slope factor and ultimate rejection.

Practical components feature undesirable characteristics, such as resistance, stray capacitance and stray inductance. For RF designs, take care to pick suitable components and minimise unnecessary detours and wiring.

Revision Questions

1 A bandpass filter:

a. allows all frequencies to pass.
b. attenuates all frequencies.
c. allows signals between two frequencies to pass.
d. increases bandwidth of a receiver.

2 A bandstop filter:

a. allows all frequencies to pass.
b. attenuates all frequencies.
c. decreases bandwidth of a receiver.
d. attenuates signals between two frequencies.

3 A lowpass filter:

a. Attenuates all signals above a known cut-off frequency.
b. Introduces harmonics.
c. Removes RF signals from an input signal.
d. Requires the use of high gain amplifiers.

4 A highpass filter:

a. Introduces harmonics.
b. Removes RF signals from an input signal.
c. Requires the use of high gain amplifiers.
d. Attenuates all signals below a known cut-off frequency.

5 A circuit which passes electrical signals above a certain frequency, but blocks electrical signals below that frequency is called:

a. an input filter.
b. a lowpass filter.
c. a highpass filter.
d. a bandpass filter.

6 The purpose of a lowpass filter is to:

a. attenuate all frequencies apart from a specific one.
b. pass all frequencies apart from a specific one.
c. pass all signals below a specified frequency but attenuate frequencies above it.
d. attenuate all signals below a specified frequency but pass frequencies above it.

7 The purpose of a highpass filter is to:

a. attenuate all frequencies apart from a specific one.
b. pass all frequencies apart from a specific one.
c. pass all signals below a specified frequency but attenuate frequencies above it.
d. attenuate all signals below a specified frequency but pass frequencies above it.

8 A lowpass filter seems to be letting through harmonic components that it is supposed to suppress. The design has been checked, and all component values are correct. The reason might well be that:

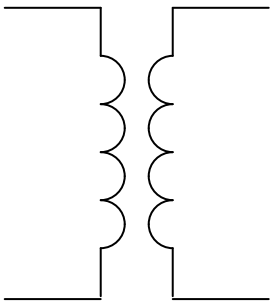
a. The components picked are not intended for RF applications.
b. The construction technique features too many interconnecting wires.
c. The components inside the cabinet are too close together.
d. Any of the above.

Chapter 14: The Transformer

14.1 Theory of Operation

The transformer is used to change (transform) the voltage and current of an AC signal. It consists of two or more windings wound on a common former. One of these windings is called the *primary winding*. The rest of the windings are known as *secondary windings*.

In operation, an AC voltage is applied to the primary winding. The resulting alternating current generates a fluctuating magnetic field, which induces a current into the secondary windings. This property is called *mutual inductance* to distinguish it from the *self inductance* that is characteristic of inductors. The schematic symbol for a transformer is shown below:



Transformer symbol

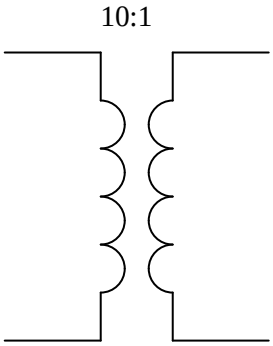
So which is the primary winding and which is the secondary? The circuit symbol does not make a distinction. The primary winding is whichever winding has power applied to it. In principle, a transformer can work both ways, depending on which side is connected to the source and which to the load.

14.2 Turns Ratio

The turns ratio of a transformer specifies the relative number of turns on the primary and secondary windings. The number of turns on the primary is always specified first. For example, a 5:1 transformer has five times as many turns on the primary as on the secondary. A 1:3 transformer has three times as many turns on the secondary as on the primary.

Note that the turns ratio does not specify the actual number of turns, just the ratio of turns on the primary with respect to the number of turns on the secondary. For example, a transformer with 200 turns on the primary and 20 turns on the secondary would be described as a 10:1 transformer because there are ten times as many turns on the primary as on the secondary. Windings could have thousands of turns.

The turns ratio is often shown in numbers on the schematic symbol, for example:



Transformer with 10:1 winding ratio

14.3 Voltage Ratio

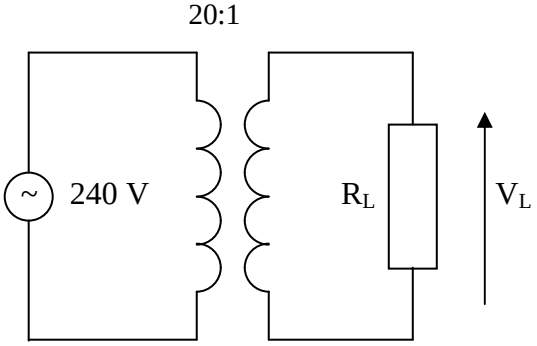
The voltages across the different windings of a transformer follow a very simple rule known as the *transformer principle*:

> **The Transformer Principle:** *The ratio of the voltage on the primary winding to the voltage on the secondary winding is the same as the ratio of the number of turns on the primary winding to the number of turns on the secondary winding (the turns ratio).*

This relationship can be written mathematically as follows:

$$V_P/V_S = N_P/N_S$$
so $$V_S = V_P N_S / N_P$$

Where N_P is the number of turns on the primary and N_S the number of turns on the secondary. For example, consider the circuit below:



Transformer circuit

240 VAC is applied across the primary of a 20:1 transformer. What is the voltage V_L across the load? From the transformer principle, the ratio of the voltage on the primary to the voltage on the secondary must be the same as the ratio of the number of turns on the primary to the number of turns on the secondary, which is 20:1. So if the voltage on the primary is 240 V, the voltage on the primary must be one twentieth of 240 V, or 12 V. Alternatively you can use the formula to get the same result:

$$V_S = V_P N_S / N_P$$
$$= 240\ V \times 1 / 20$$
$$= 12\ V$$

A transformer with more turns on the secondary than on the primary will have a higher voltage across the secondary than across the primary, and is called a *step-up* transformer because it "steps the primary voltage up" to a higher secondary voltage. A transformer with fewer turns on the secondary than on the primary will have a lower voltage across the secondary than across the primary and is called a *step-down* transformer. For example, a 1:5 transformer is a step-up transformer, while a 10:1 transformer is a step-down transformer.

14.4 Current Ratio

For a transformer with a single secondary winding, the ratio of the current in the secondary to the current in the primary is the same as the ratio of the number of turns in the primary to the number of turns in the secondary (the turns ratio). Note that this is the opposite way round to the voltage ratio, so a step-up transformer (which has a greater voltage across the secondary than across the primary) will have a smaller current flowing in the secondary than in the primary; while a step-down transformer (which has a smaller voltage across the secondary than across the primary) will have a larger current flowing in the secondary than in the primary.

Remember that the total power coming out of the transformer must the same as the power flowing into it (neglecting minor losses). Therefore, if the voltage increases, the current must decrease to maintain the same total power. Likewise, if the output voltage is lower, the output current must be higher to maintain the same power.

Mathematically this relationship can be expressed as follows:

$$I_S/I_P = N_P/N_S$$

or

$$I_S = I_P\, N_P/N_S$$

So for the example above, suppose the current flowing in the primary is 1 A. Then the current in the secondary can be found from

$$I_S = I_P\, N_P / N_S$$
$$= 1 \times 20 / 1\ A$$
$$= 20\ A$$

The power drawn by the primary is:

$$P_P = V_P\, I_P$$
$$= 240 \times 1\ W$$
$$= 240\ W$$

The power supplied to the load by the secondary is:

$$P_S = V_S\, I_S$$
$$= 12 \times 20\ W$$
$$= 240\ W$$

Since the power supplied to the load by the secondary is identical to the power drawn by the primary, the transformer has not dissipated any power. In practical transformers there is usually some small power dissipation caused by the resistance of the windings and eddy currents flowing in the transformer core. Typical power transformers are better than 95% efficient, which means that less than 5% of the power is lost in the transformer.

The ability of transformers to efficiently and simply convert high voltages to low voltages and *vice versa* is the principal reason why the mains supply in all countries is AC. The distribution network uses voltages of up to 800 kV with relatively low current. Low current

minimises heating losses in transmission lines. For safety reasons, consumers must be provided with lower voltage. The closer the power gets to the end user, the lower the voltages become. Local distribution systems use voltages of around 11 kV, and the final voltage delivered to the customer is normally around 240 V.

14.5 Impedance Ratio

We haven't quite finished with our example circuit yet. Since we know the voltage across the load resistance and the current flowing through the load, we can calculate the resistance:

$$\begin{aligned} R_L &= V_S / I_S \\ &= 12 / 20 \\ &= 0{,}6\ \Omega \end{aligned}$$

We can also calculate the resistance that the primary winding appears to have to the voltage source driving it.

$$\begin{aligned} R_P &= V_P / I_P \\ &= 240 / 1 \\ &= 240\ \Omega \end{aligned}$$

Note that this "resistance" is not the actual resistance of the primary winding, which would typically have a resistance of less than 1 Ω. It is rather an apparent resistance caused by the fact that power is being drawn from the primary winding. In this case though the power is ending up in the secondary circuit and is not being dissipated by the transformer itself but rather by the load.

Another way to look at it is that the voltage source driving the primary is "seeing" the load resistance, but the transformer has transformed the actual value of the resistance, just as it has transformed the voltage and current values. We can derive a general rule for this resistance transformation:

$$\begin{aligned} R_L &= V_S / I_S \\ &= (V_P\ N_S / N_P) / (I_P\ N_P / N_S) \\ &= (V_P / I_P)\ (N_S^2 / N_P^2) \\ &= R_P\ N_S^2 / N_P^2 \\ &= R_P\ (N_S / N_P)^2 \end{aligned}$$

conversely,

$$R_P = R_L\ (N_P / N_S)^2$$

where R_P is the "apparent resistance" of the primary winding. The fact that the load resistance in the secondary circuit causes a different (but related) resistance to appear in the primary circuit is known as "impedance transformation". Impedance is a general concept that combines resistance and reactance, which is covered in more detail in another module.

These equations tell us that the resistance in the primary circuit and the resistance in the secondary circuit are related by the *square* of the turns ratio. The resistance transformation works in the "same direction" as the voltage transformation so for a step-down transformer, where the voltage across the secondary is smaller than the voltage across the primary, the resistance in the secondary circuit will also be smaller than the resistance in the primary circuit. Similarly, a step-up transformer will "step up" the resistance in the primary circuit to a larger resistance in the secondary circuit. However, note that the amount by which impedances are transformed is not the same as the amount by which voltages are

transformed since the impedance transformation depends on the *square* of the turns ratio, while the voltage transformation depends on the turns ratio (not squared).

For example, suppose you have an audio amplifier designed to drive a 200 Ω load but you want to connect it to an 8 Ω speaker instead. You could use a transformer to convert the impedances. You need a 200:8 or 25:1 impedance transformation from the primary (connected to the amplifier output) to the secondary (connected to the speaker). A 5:1 transformer could provide this transformation. Note that the turns ratio, being 5:1, is the *square root* of the impedance ratio, which is 25:1.

14.6 Applications

Transformers are widely used in amateur radio. The most obvious example is in old-style power supplies, where the mains voltage of 240 V must be transformed to a voltage suitable for running radio equipment, typically 12 V DC.

Transformers are also widely used for *impedance matching* within transmitter and receiver circuits. For example, an antenna system typically has an impedance of 50 Ω, while the RF amplifier of a typical receiver would generally have a much higher input impedance. Since maximum power is transferred when the source and load impedances are equal, a transformer might be used to match the impedance of the antenna to the input impedance of the RF amplifier.

Summary

An AC voltage applied to the *primary winding* of a transformer generates a fluctuating magnetic field that induces a voltage in the *secondary winding*. This phenomenon is known as *mutual inductance*.

The ratio of the voltage on the primary winding to the voltage on the secondary winding is the same as the ratio of the number of turns on the primary winding to the number of turns on the secondary winding (the *turns ratio*).

$$V_S = V_P N_S / N_P$$

A transformer with more turns on the secondary than on the primary is a *step-up* transformer, which would increase the applied voltage; one with fewer turns on the secondary than on the primary is a *step-down* transformer, which decreases the applied voltage.

The ratio of the current in the secondary to the current in the primary is the inverse of the turns ratio:

$$I_S = I_P N_P / N_S$$

The overall effect is that the power in the primary circuit is equal to the power in the secondary circuit, so a perfect transformer does not dissipate power. Practical transformers exhibit losses of perhaps 5%.

By transforming voltages and currents, transformers also transform the load resistance to an apparent resistance in the primary winding. The transformation occurs in the same “direction” as the voltage transformation, but according to the *square* of the turns ratio:

$$R_P = R_L (N_P / N_S)^2$$

Revision Questions

1 The principle of operation of a transformer is based upon:

a. Static electricity.
b. Potential difference.
c. Electrostatics.
d. Electromagnetic induction.

2 Transformers transfer energy from one coil to another by means of:

a. Inductive coupling.
b. Static discharge.
c. Capacitance.
d. Electrical conduction.

3 A transformer with a turns ratio of 1:8 is called a:

a. step down transformer.
b. step up transformer.
c. low current transformer.
d. high-tension transformer.

4 A transformer nameplate shows a figure of 1:4. If 12 V AC is applied to the primary winding, what is the voltage on the secondary terminals?

a. 3 V
b. 48 V
c. 16 V
d. 8 V

5 Losses in transformers are most often caused by winding resistance and:

a. Stray capacitance.
b. Mutual inductance.
c. Eddy currents.
d. Loss leaders.

6 What is the turns ratio of a transformer to match an audio amplifier having an output impedance of 200 Ω to a speaker having a load impedance of 10 Ω?

a. 4,47:1
b. 14,14:1
c. 20:1
d. 400:1

7 The operating principle of a transformer may be described as:

a. A varying magnetic field intersecting a conductor and creating a potential difference.
b. A varying electric field intersecting a conductor and creating a potential difference.
c. A varying current in a conductor setting up a static magnetic field.
d. A varying voltage in a conductor setting up a static magnetic field.

8 An impedance-matching transformer has a turns ratio of 10:1. If a 500 Ω microphone is connected to the winding with the lesser turns, it would correctly operate into a load of:

a. 5 Ω
b. 50 Ω
c. 50 kΩ
d. 500 kΩ

9 A transformer has 1200 turns on its primary and 30 turns on its secondary. If it is connected to the mains supply (240 V), the secondary voltage will be:

a. 9,6 kV
b. 240 V
c. 30 V
d. 6 V

10 A 5:1 transformer has a current of 1 A flowing in the primary winding. The current flowing in the secondary winding is:

a. 40 mA
b. 200 mA
c. 5 A
d. 25 A

Chapter 15: Semiconductors and the Diode

15.1 Semiconductors

Semiconductors are materials where the outer electrons are more tightly bound to the nucleus than in the case of conductors, but less tightly bound than in the case of insulators. Normally at room temperature these materials behave as insulators; but when heated, the additional energy of the electrons allows the outer ones to break away from the nucleus, so allowing a current to flow if an electric potential is applied.

The semiconductor most commonly used in electronic devices is silicon. Silicon atoms have 14 electrons, arranged in three layers. The first (inner) layer has 2 electrons; the second layer has 8 electrons; and the outer layer has 4 electrons. It is the electrons in this outer layer that are only moderately bound to the nucleus. Silicon atoms normally form a crystal lattice with other silicon atoms, where each atom shares one of its outer electrons with another atom in the lattice.

When silicon is used to manufacture electronic components, it is first refined to make it very pure silicon, and then small quantities of another material are introduced. This process is known as “doping”.

Early semiconductor devices were mostly made from germanium. These days most devices are made from silicon, but special-purpose semiconductors are made from other materials such as gallium arsenide (GaAs).

N-Type Semiconductors

Suppose a very small quantity of a material with five electrons in its outer shell such as phosphorous or arsenic is added to very pure molten silicon and then the mixture is allowed to cool and crystallise. The deliberate impurity is know as a doping agent. The silicon atoms will take up their normal crystal structure, with each atom sharing one electron with each of its four neighbouring atoms. The occasional doping atom will be forced to fit in with this structure, so it will also share one of its outer electrons with each of its four neighbouring silicon atoms. However, since the doping agents have *five* outer electrons, one free electron is not bound into the crystal structure. This electron will be free to migrate around the crystal lattice, and can serve as a charge carrier, allowing a current to flow if a potential is applied. Silicon doped in this way becomes an electrical conductor at room temperature, due to the free charge carriers. Since the charge carriers are negatively charged electrons, this semiconductor is called an “N-type”, where the N stands for “negative”.

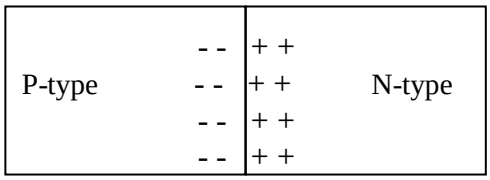
P-Type Semiconductors

Now suppose that we add a doping agent with only three outer electrons, such as boron or aluminium, to pure molten silicon and allow it to cool and form a crystal lattice. Again the silicon atoms will take up their normal crystal structure, with each atom sharing one electron with each of its four neighbours. The occasional doping atom will be forced to fit into the structure, sharing one of its outer electrons with each of its four neighbouring silicon atoms. However, since the doping agent only has *three* outer electrons, one of its neighbours will have to do without a shared electron. This shortage leaves a “hole” in the crystal lattice, a place where there ought to be an electron but there isn’t one. If an electric potential is applied across such a material, the electrons will be attracted by the positive terminal and repelled by the negative terminal. However, most of the electrons will be unable to move as they are rigidly bound into the lattice structure. However, the electron that is on the negative potential side of the “hole” can move to the place in the lattice where the electron is missing, leaving its own place in the lattice structure empty. Another electron can fill this empty slot, leaving its own place empty, and so on. In this way the “hole” can appear to migrate across the lattice, even though electrons are actually moving. Because the

hole is the absence of a negatively charged electron, it behaves as though it was a positive charge that is free to move around the lattice, For example, if a potential difference is applied, holes will migrate from the positive terminal to the negative terminal. In this way, holes can also act as charge carriers, and since they behave like positively charged charge carriers, semiconductors doped in this way are known as "P-type" semiconductors.

15.2 The Junction Diode

Now suppose that a small piece of P-type semiconductor is brought into contact with a small piece of N-type semiconductor. The contact area between them is called a "PN junction".



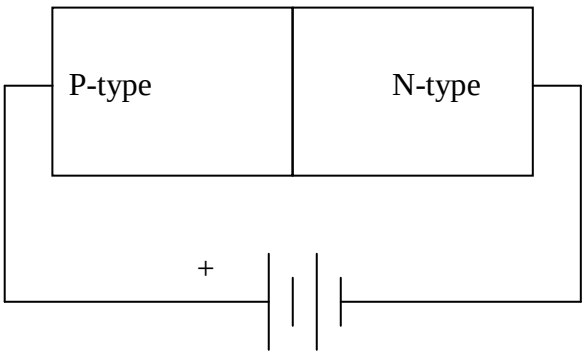
The PN Junction

Because the N-type material has some free electrons, some of these can migrate across the boundary and "fill" some of the holes in the lattice structure of the P-type material. This will leave the P-type material negatively charged, while the N-type material will become positively charged. The process will stop when the potential difference between the P-type and N-type materials is sufficient to prevent any further electron movement across the boundary.

Remember that the P and N type materials were *not* positively and negatively charged to start with. They were both neutral since the "extra" electrons in the N-type material were balanced by the additional positive charges on the nuclei of the phosphorous or arsenic atoms used for doping. Similarly, the "lack" of electrons in the P-type material is precisely balanced by the smaller positive charge on the nuclei of the boron or aluminum atoms.

A very thin layer called the *depletion layer* or *depletion region* has now been formed at the junction between the P-type and N-type materials where there are no (or very few) free charge carriers since the free electrons from the N-type material have all been "used up" filling the holes on the P-type side of the junction. The depletion layer is usually very thin, only 1 µm or so.

Now suppose we apply a potential difference across the junction, with the positive terminal attached to the P-type material and the negative terminal to the N-type material.

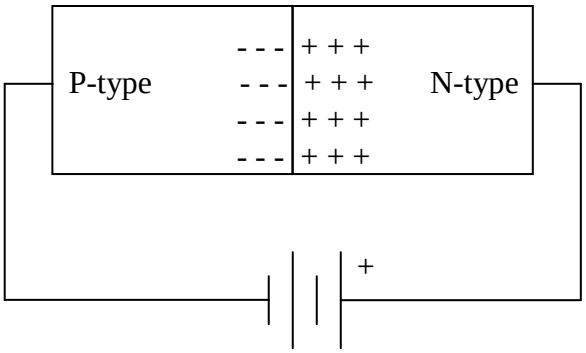


A Forward-Biased PN Junction

The battery will effectively "pump" electrons into the N-Type material and towards the depletion region. Similarly, it the positive potential applied to the P-type material will attract electrons away from the depletion region. The net effect is that there now are charge carriers in the depletion region—electrons in the N-Type material, and holes in the P-type material—and so a current can flow. Another way of looking at it is that the depletion layer has been neutralised by the application of a potential difference across the forward-biased junction.

In order to make this current flow, there must be sufficient potential difference to overcome the potential difference that existed across the junction between the N-type and P-type materials due to the migration of electrons from the N-type material to the P-type material when the depletion layer was formed. This voltage is known as the "forward bias voltage" and is typically between 500 and 800 mV in silicon PN junctions and around 100 to 200 mV for germanium junctions. If the potential difference applied is less than this then electrons won't be forced into the depletion region on the N-type side, or removed from it on the P-type side, so the depletion region will still not have any free charge carriers and no current will flow.

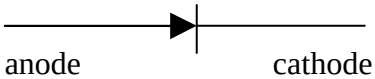
Now let's see what happens if we connect the potential the other way around, with the positive terminal attached to the N-type material and the negative terminal to the P-type material.



A Reverse-Biased PN Junction

The effect of the applied potential is to force more electrons into the P-type material, making it more negatively charged, and remove electrons from the N-type material, making it more positively charged. This movement of charges simply results in more holes in the P-type material being "filled" by the additional electrons, and more of the free electrons being removed from the N-type material, increasing the depth of the depletion layer and increasing the potential difference across the junction until it equals the potential difference applied from outside. So after a very brief initial current, no further current will flow except a tiny current known as the "reverse leakage" current which is typically less than 1 μA. This situation is known a "reverse-biased" junction.

The device we have described from a physical point of view is called the *junction diode*, often just "diode" for short. Its circuit symbol is shown below:



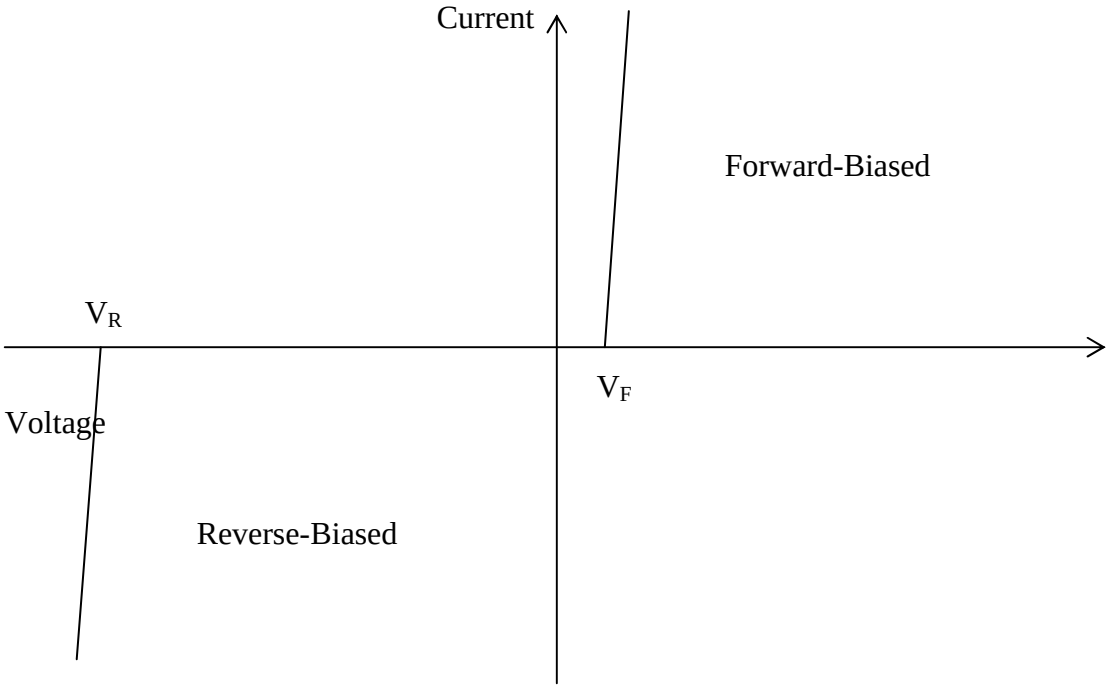
A semiconductor diode

The two terminals of the diode are called the "anode" and the "cathode". The diode will allow a current to flow if the anode is more positive than the cathode by at least the amount

of the forward bias voltage (200 to 800 mV, as discussed). This current flows in the direction of the arrow, i.e. from left to right in the diagram above.

If the diode is reverse-biased, only a tiny current, the reverse leakage current, will flow. Of course if you apply a high enough potential across a reverse-biased junction then eventually the depletion layer will break down, allowing a current to flow. However, in most diodes this breakdown is likely to cause permanent damage to the diode. There is an exception; the Zener diode, which is discussed below.

The graph below plots the voltage across the diode against the current flowing through the diode, using the convention that a positive voltage means the diode is forward-biased, and a negative voltage means it is reverse-biased.

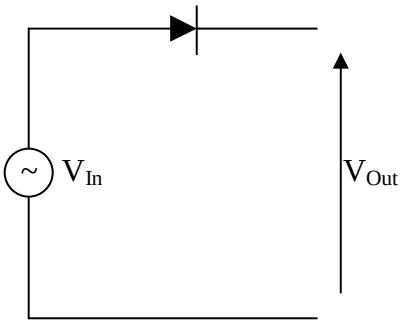


Voltage vs Current for a semiconductor diode

When the diode is forward-biased, no current flows until the voltage applied exceeds the forward-biased voltage of the diode, V_F. Once the voltage exceeds V_F, the current rises rapidly with little change in the voltage across the diode. For this reason, it is a good approximation to assume that *the voltage across a forward-biased diode is always V_F, the forward bias voltage*.

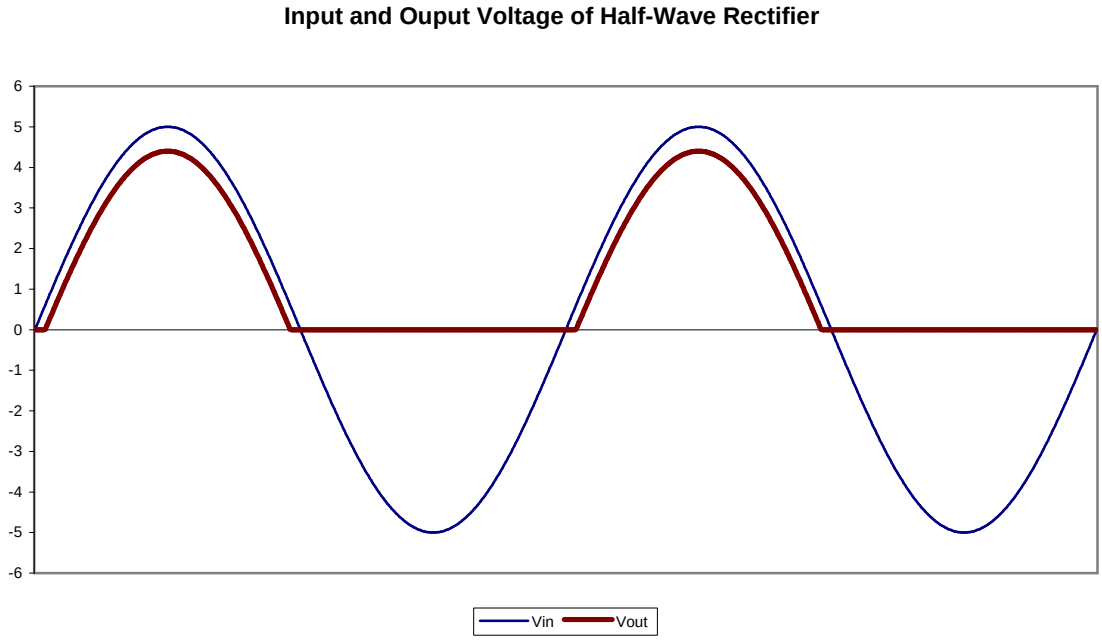
15.3 The Half-Wave Rectifier

Diodes are commonly used as *rectifiers* to convert AC voltages into DC voltages. For example, consider the circuit below, which is known as a *half-wave rectifier*.



AC source with half-wave rectifier

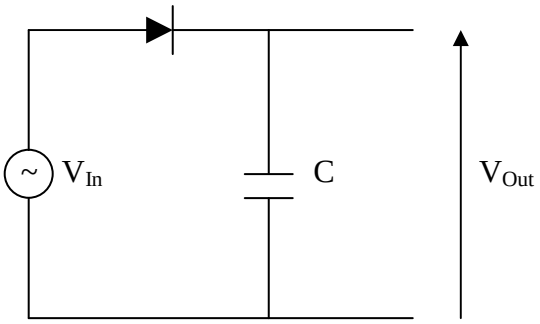
The graph below shows the input and output voltages for a half-wave rectifier:



Input and output for half-wave rectifier with sinusoidal input

The output voltage follows the input voltage during the positive half cycles, but is always slightly less than the input voltage due to the forward-bias voltage drop (in this case 600 mV) across the diode. During the negative half cycles, the diode does not conduct and the output voltage is zero.

Of course the resulting voltage can hardly be considered “DC” since it is still varying periodically. However, this problem can be solved using a *smoothing capacitor*.



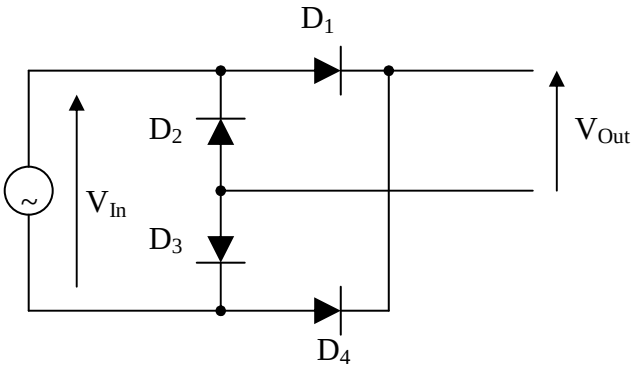
Half-wave rectifier with smoothing capacitor

The "smoothing capacitor" C charges up during the positive half cycle, and then discharges into the load (which is not shown) during the negative half cycle, keeping a relatively constant DC voltage across the load. However, some traces of the original alternating voltage will remain superimposed on the DC output voltage. This variation is known as "ripple" and can cause problems such as hum in audio amplifiers. For a half-wave rectifier, the ripple has the same frequency of the AC signal that is rectified, most commonly 50 Hz in South Africa.

Another way of thinking of this circuit is that the capacitor C forms a simple one-element lowpass filter. The half-wave rectified sine signal contains both a DC component and an AC component, and the capacitor attenuates the AC ripple component while passing the DC output.

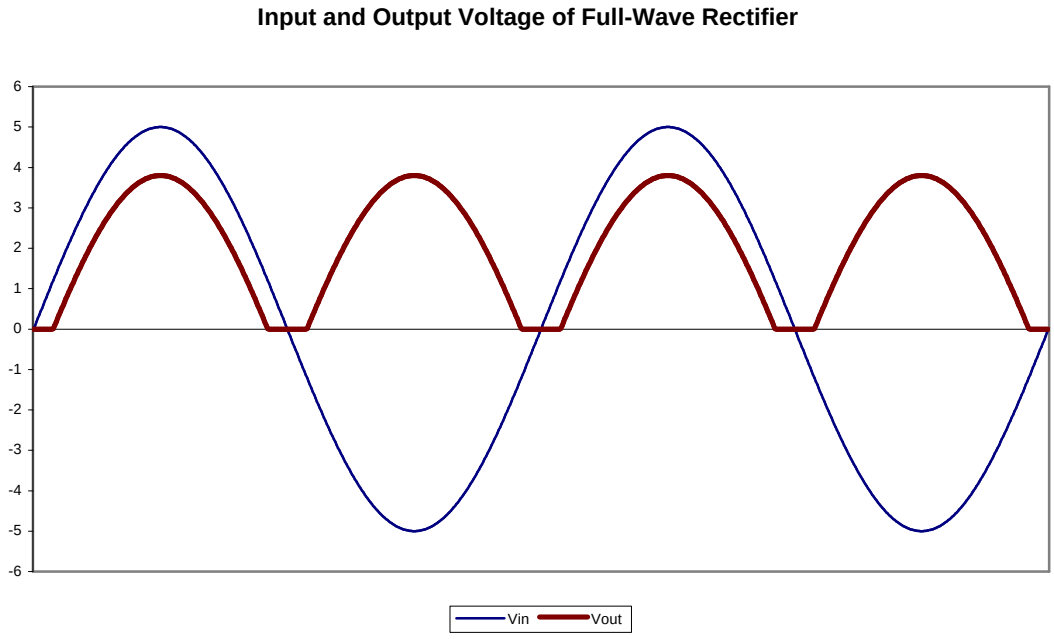
15.4 The Full-Wave Rectifier

Half-wave rectifiers suffer from excessive ripple because they give zero voltage output for just over half of each cycle. The full-wave rectifier shown below is better in this regard.



Full-wave rectifier circuit

When V_{In} is positive, D_1 and D_3 conduct and V_{Out} is positive. When V_{In} is negative, D_2 and D_4 conduct and V_{Out} is still positive! This circuit is known as a *full-wave bridge rectifier* and the graph below shows the input and output voltages.



Input and output waveforms for full-wave rectifier with sinusoidal input

The output voltage is now positive during both the positive and negative half-cycles of the input voltage. Note that the output voltage is always about 1,2 V less than the input voltage, because there are now *two* 600 mV forward voltage drops across the two conducting diodes. Note also that the frequency of the "ripple" on the output is now *twice* the input frequency (mostly 100 Hz in South Africa).

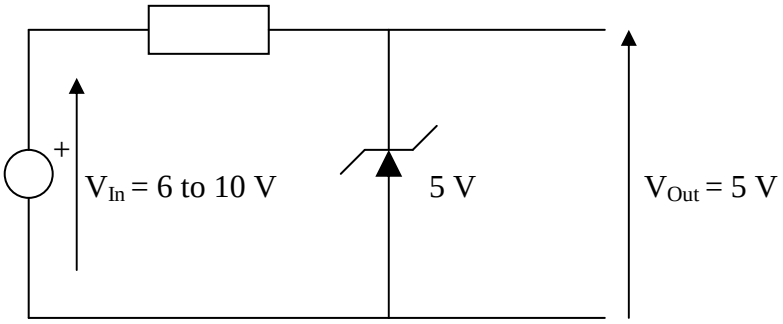
Once again, the amount of ripple on the output can be greatly reduced by using a suitable smoothing capacitor as a lowpass filter on the output. The capacitor can generally be much smaller than for a half-wave rectifier, as the discharge time is less than half as long in this case.

15.5 Special Diodes

The Zener Diode

In most diodes, the reverse breakdown voltage V_R should never be exceeded or the diode may be permanently damaged. However, *Zener diodes* are designed so that the reverse breakdown voltage can be safely exceeded without damage to the diode. Zener diodes are also manufactured in such a way that the reverse breakdown voltage (also known as the *Zener voltage* V_Z) is specified and carefully controlled. Zener diodes are available with Zener voltages ranging between 2 V and 100 V.

Zener diodes are commonly used as voltage regulators. For example, consider the circuit below:



Zener diode regulator with series resistor

This diagram shows an input voltage in the range 6 to 10 V being applied across a *reverse-biased* Zener diode with a Zener voltage of 5 V, through a resistor. The purpose of the resistor is to limit the current flowing through the Zener diode and prevent it from being destroyed. Note the circuit symbol for the Zener diode.

If the output voltage rises above 5 V, the Zener diode conducts strongly even though it is reverse-biased. The current flowing through the Zener diode causes a voltage drop across the resistor, reducing the output voltage until it is close to 5 V. In this way, the circuit maintains a fairly constant output voltage of around 5 V irrespective of both the input voltage and the current drawn by the load. The reverse-biased Zener diode functions as a *voltage regulator*, maintaining a constant output voltage despite fluctuations in the input voltage and load current.

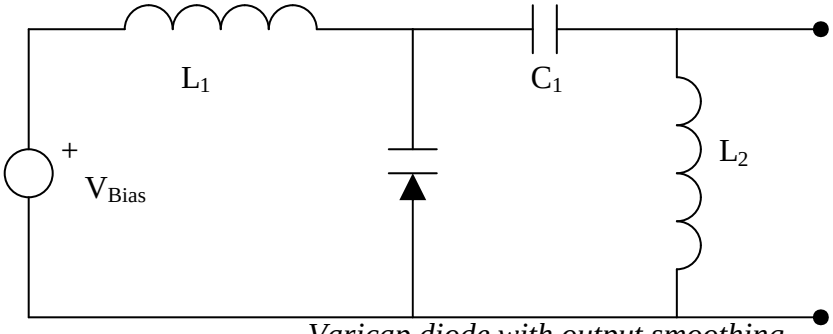
The Varicap Diode

Think back to the physical description of the reverse-biased PN junction. It consists of two conducting materials (P-type and N-type semiconductors) separated by the thin non-conducting depletion layer.

This description is very similar to the description of a capacitor: two conducting plates separated by a thin insulating layer. Indeed, all reverse-biased diodes do act as capacitors to some extent. Most diodes are designed to minimise the capacitance effect so that it becomes negligible except at very high frequencies.

However, some diodes are designed to maximise this capacitance effect and to allow the capacitance to be controlled by the reverse-bias voltage applied to the diode. Remember that the greater the reverse-bias voltage, the larger the depletion layer. This change in layer thickness is equivalent to moving the plates of a capacitor further apart—in other words, the capacitance will be reduced. These are called varicap ("variable capacitance") diodes. An alternative term is a *varactor diode*.

A typical circuit is shown below:



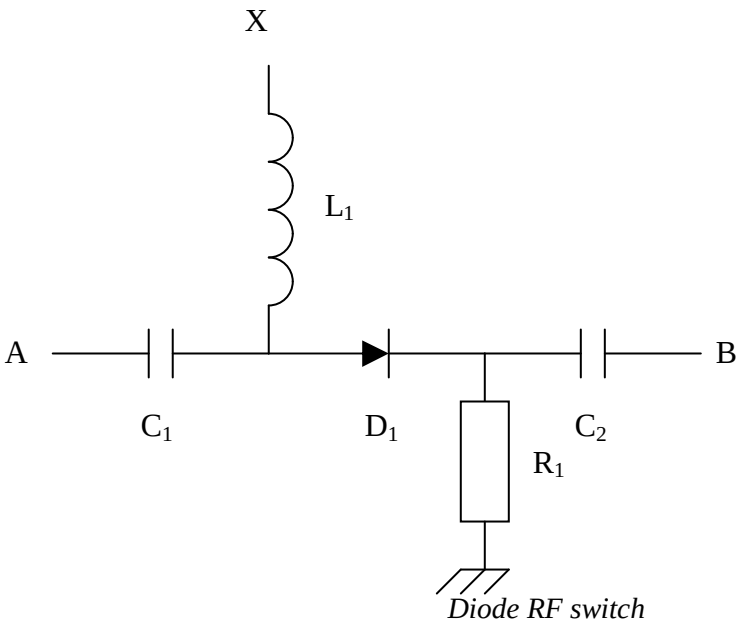
Varicap diode with output smoothing

As you can see, the circuit symbol for a varicap diode is appropriately a combination of the symbols for a diode and for a capacitor.

In this circuit, a DC bias voltage V_{Bias} is used to reverse-bias varicap diode D_1. As the bias voltage is changed, the capacitance of D_1 is changed. This change in capacitance will vary the resonant frequency of the parallel tuned circuit made up of L_2, C_1 and D_1. Note that the capacitance in the parallel tuned circuit consists of C_1 in series with the varicap diode. L_1 prevents the AC signals present in the tuned circuit from flowing through the bias voltage source, which would introduce undesirable losses into the tuned circuit, reducing its Q. Similarly C_1 prevents the DC bias voltage from being shorted through L_2. Circuits like this are often used to vary the frequency of oscillators (that is, circuits which generate AC signals at a specific frequency) in response to a DC tuning voltage.

Diodes as Switches

Although it may appear surprising at first, diodes are often used as switches in radio frequency (RF) applications.



Diode RF switch

Suppose an AC voltage is applied at point A. Will it reach B? Let us first assume that the control voltage X is positive with respect to the chassis. The DC current will flow through inductor L_1 and diode D_1, and to the chassis via resistor R_1. The conducting metal chassis is represented by the symbol below the resistor. Because the diode is forward-biased, it offers very little resistance to the flow of current, so a signal applied at point A is coupled through capacitor C_1, diode D_1 and capacitor C_2 to point B. As long as the signal applied at A is significantly smaller than the control voltage applied at X, the signal will *not* be rectified since the diode will be forward-biased even during the negative peaks of the signal.

Now consider what happens if a negative voltage (with respect to the chassis) is applied to X. Now D_1 is reverse-biased, and will not conduct current. Again assuming that the signal applied at A is significantly smaller than the control voltage, the diode will remain reverse-biased even at the positive peaks of the signal. Hence the signal will effectively be blocked by the reverse-biased diode.

The role of L_1 is simply to have high reactance at the signal frequency, preventing the signal from being coupled into the control circuitry, while allowing the DC control current to flow unimpeded. The capacitors do exactly the opposite; they block DC from reaching

points A and B, while allowing the RF signal to flow through them unimpeded. Such capacitors are known as *decoupling capacitors*.

Diodes are widely used in this way in modern microprocessor-controlled transceivers since the microprocessor can easily generate suitable control voltages, which can be used to perform various RF switching tasks, such as switching between different filters. The best diodes for switching tasks are PIN diodes, which have a low resistance when forward-biased, and high resistance when reverse-biased, combined with low capacitance. Significant capacitance across the reverse-biased diode would be disastrous in this circuit, as it would allow the signal through even when the switch was turned "off"! PIN diodes contain some pure ("intrinsic") silicon between the N and P regions—hence the PIN.

PIN diode switches have advantages over mechanical switches such as relays. They do not have an inherent lifetime limit—they can survive a very high number of cycles. They can also switch much faster than typical mechanical devices.

The Light-Emitting Diode

If the semiconductor material and doping are appropriately chosen, the energy state transition around the junction can emit electromagnetic energy. Such a diode is known as a *light-emitting diode* (LED).

LEDs are available in different wavelengths, from infrared to visible to ultraviolet. The first LEDs were red, green and yellow, but nowadays there is a full complement of LEDs. Sophisticated displays, e.g. for TV sets, can be made by combining LEDs in the three primary emission colours (RGB for red, green and blue) into each pixel. If all three are turned on, a white colour is emitted. With different combinations of intensities, virtually any colour can be reproduced.

LEDs offer high reliability, degrade very slowly and are much more efficient than incandescent or fluorescent lights. Replacing an existing lamp with LEDs provides lower power consumption and more light. However, some users feel that the light emitted by an LED lacks "warmth". This objection is normally because users are used to lots of yellow and red light, such as they obtained from old globes. LED lighting is often rather blue, with little red or yellow component.

Most LEDs have a relatively high forward voltage—typically in the range of 2,0 V. Current requirements range from 2 mA to perhaps 200 mA for high-intensity LEDs.

Summary

Pure semiconductors do not conduct current at room temperature. However, by introducing small amounts of impurities they can be made to conduct a current. The charge carriers in N-type semiconductors are negatively charged electrons; while in P-type semiconductors they are positively charged holes.

When N-type and P-type semiconductors are brought into contact, a thin *depletion layer* with no free charge carriers forms at the junction. If the junction is *forward-biased* by making the P-type material positive with respect to the N-type layer, a current will flow provided the potential exceeds the forward bias voltage of the junction, which is typically around 600 mV for silicon devices. If the junction is *reverse-biased* by making the P-type material negative with respect to the N-type material then no current will flow until the *reverse breakdown* voltage is reached.

The PN junction is used to make an electronic device called the *junction diode*. The diode has two terminals: the anode, which is connected internally to the P-type material; and the cathode, which is connected to the N-type material. If the anode is made positive with

respect to the cathode by 600 mV or so (the forward bias voltage) then a current will flow through the diode.

Diodes are commonly used as rectifiers. In a *half-wave rectifier*, only the positive half cycle is rectified, so there is a substantial amount of *ripple* at the AC frequency, which must be removed using a *smoothing capacitor*. In a *full-wave rectifier* both the positive and the negative half-cycles are rectified, resulting in less ripple. In a full-wave rectifier the ripple frequency is *twice* the frequency of the AC input. In audio systems, this ripple is known as *hum*.

Zener diodes have a well-controlled and specified reverse breakdown voltage, also known as the *Zener voltage*, and are designed not to be damaged by reverse breakdown. When reverse-biased, Zener diodes act as voltage regulators. *Varicap diodes* exhibit a capacitance when reverse-biased that decreases as the reverse bias voltage is increased. *PIN diodes* are often used as switches at radio frequencies.

Light-Emitting Diodes (LEDs) provide light in virtually any colour, very efficiently. The forward drop of an LED is typically much higher than for simple junction diodes.

Revision Questions

1 The forward-biased diode junction will:

a. Allow current to flow through the junction.
b. Block current from flowing.
c. Have a high resistance.
d. Exhibit a zener diode function.

2 In a silicon diode the voltage of 600 mV refers to:

a. The breakdown voltage.
b. The forward bias voltage.
c. The zener voltage.
d. The cut-off voltage.

3 Which of the following components is intended only to operate in the reverse-biased condition?

a. A rectifier diode.
b. A zener diode.
c. A polarised capacitor.
d. A resistor.

4 The term Vz is commonly used to describe:

a. The zener diode regulating voltage.
b. The zener diode impedance.
c. The forward voltage applied to the diode.
d. The peak voltage of the rectified waveform.

5 By only allowing alternating current to flow in one direction a diode can be used as:

a. An attenuator.
b. An amplifier.
c. A rectifier.
d. A fuse.

6 Zener diodes are used to:

a. Start oscillations.
b. Divert signals.
c. Detect modulation.
d. Regulate a DC voltage supply.

7 What function does a full-wave bridge circuit serve?

a. Amplification.
b. Coupling.
c. Rectification.
d. Isolation.

8 A circuit which only allows half of an AC waveform to pass through is called:

a. A regulator.
b. A bridge circuit.
c. An attenuator.
d. A half-wave rectifier.

9 A four-diode circuit to produce full-wave rectified DC from a transformer is called:

a. A balanced circuit.
b. A bridge rectifier.
c. A dummy load.
d. A regulator.

10 The area of a diode junction where no free holes or electrons exist is called the:

a. Anode.
b. Cathode.
c. Depletion region.
d. Semiconductor.

11 An PN type semiconductor refers to a:

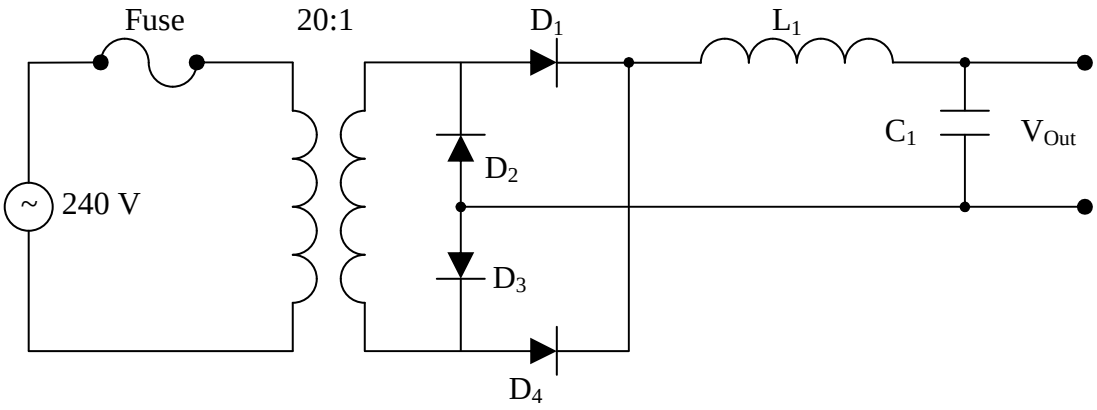
a. Two pin transistor.
b. A capacitor.
c. A diode.
d. A power resistor.

Chapter 16: The Power Supply

16.1 Simple Power Supply

A power supply is the somewhat misleading term for a device that turns the mains AC supply into a DC voltage for use with electronic equipment.

The circuit below shows a simple unregulated DC power supply that is easy to construct.



Simple unregulated DC power supply

You should recognise the various parts of the circuit and understand each part's function. The power supply applies 240 V AC mains input to the primary winding of a 20:1 transformer through a fuse. The fuse is there to protect the circuit if too much current is drawn, either by overloading the output or due to a circuit fault. The transformer steps the voltage down to 12 V AC. This voltage is rectified by a full-wave bridge rectifier consisting of diodes D_1 to D_4. The resulting waveform is passed through a lowpass filter consisting of inductor L_1 and smoothing capacitor C_1, which reduce the ripple to an acceptably low level.

The inductor also provides "inrush protection", preventing the transformer from being damaged by the high current that would flow when the supply was first turned on, when the capacitor is completely discharged. Instead of allowing a very high current to flow initially, the inductor opposes the change in current, allowing it to build up more gradually. In practice, this inductor is often omitted in simple designs, with the self-inductance of the transformer secondary being sufficient to prevent damage.

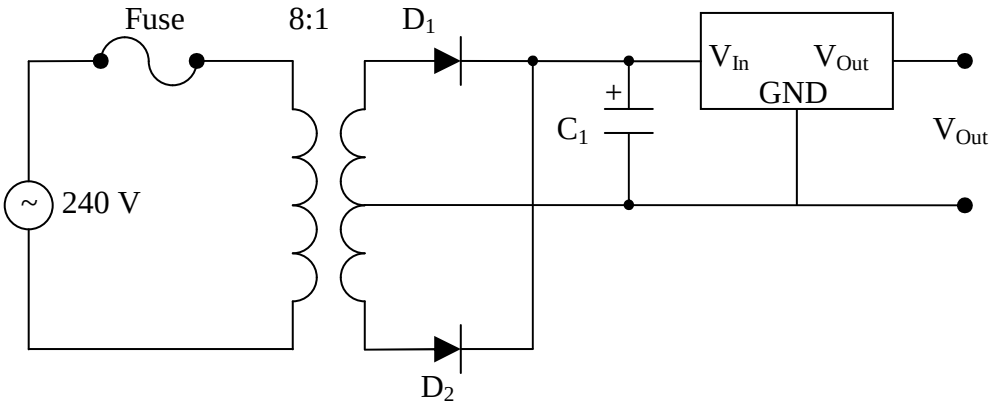
V_{OUT} is actually higher than 12 V. Remember that AC voltages are normally specified as RMS, which is about 71% of the peak voltage for sinusoidal waveforms. The peak voltage is some 41% higher than RMS, in this case about 17 V.

16.2 A Regulated Power Supply

Although the simple power supply is quite practical, it has two weaknesses. First, although the ripple is significantly attenuated by the lowpass filter on the output, some ripple will remain, which may cause problems with sensitive equipment. Second, although the output voltage is nominally 17 V, in practise it will vary between about 11 and 17 V depending on the circuit load, which again may cause problems for sensitive equipment.

Both of these problems can be solved by adding a voltage regulator to the basic design. Although a Zener diode could be used as a voltage regulator, they are only suitable for low-current applications, so they are typically used to stabilise a reference voltage that is then used to control the output voltage of the voltage regulator. Although the entire regulator can be (and often is) made from discrete components like diodes, capacitors and transistors, we will use an integrated circuit that is specifically designed as a voltage regulator. An

integrated circuit consists of a number of different electronic devices all fabricated (made) and interconnected together on a single wafer of silicon. They are available to perform many common tasks, including voltage regulation.



Simple regulated DC power supply

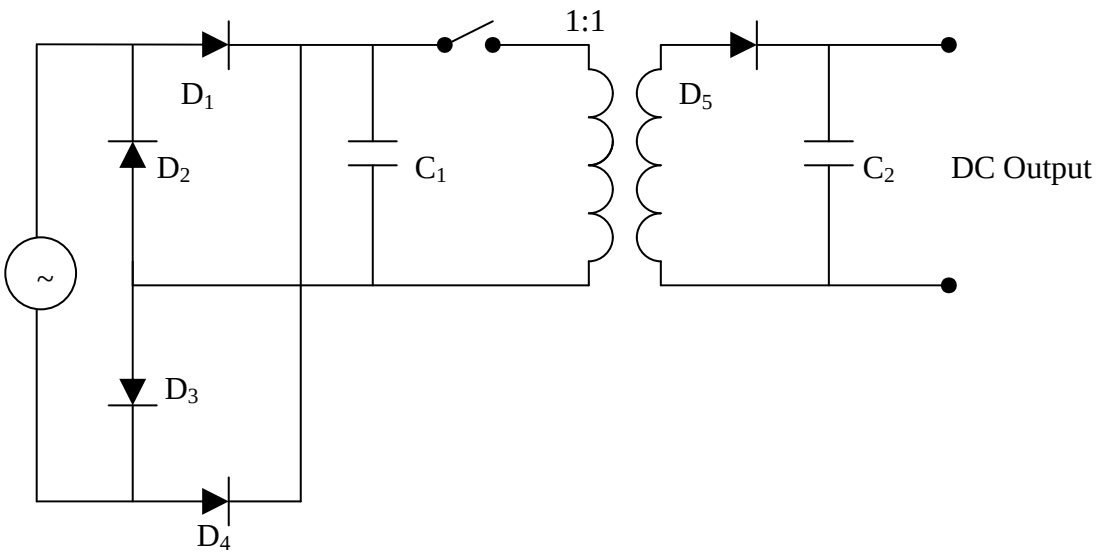
You will notice that he rectifier design is different. Instead of using four diodes in a bridge circuit, this design only uses two diodes. However, it still achieves full-wave rectification by making use of a *centre tap* on the secondary winding of the transformer. This is just a separate connection to the middle of the secondary winding. This layout allows the secondary to function almost like two separate windings. On each half cycle, either D_1 or D_2 will conduct, but not both. Whichever diode conducts will connect the positive side of the transformer to the positive side of C_1. Current flowing back to the centre tap of the secondary completes the circuit. In effect, only half of the secondary winding is used in each half cycle.

The voltage regulator has three terminals labelled V_{In}, V_{Out} and GND (for "ground"). It acts a bit like a variable resistor that is connected between the V_{IN} and V_{OUT} terminals and which is continually adjusted to maintain the voltage between the V_{OUT} and GND terminals at a constant level, say 14 V. As well as regulating the voltage, the voltage regulator also substantially reduces the ripple, since it is able to change its internal resistance fast enough to counteract the voltage fluctuations due to the ripple, thus maintaining a good "clean" DC supply despite considerable ripple in the input voltage. It is therefore possible to simplify the lowpass filter by removing the inductor, leaving only the smoothing capacitor C_1, which should provide adequate ripple rejection when used with a voltage regulator integrated circuit.

Most integrated voltage regulators provide another benefit: they automatically limit the current and power dissipation of the regulator to safe limits, avoiding damage to the power supply even if the load is short circuited. However, the mains supply of a power supply (or any other mains-powered device) should always be fused, just in case a short circuit comes about within the power supply itself.

16.3 Switching Power Supplies

Power supplies that use a transformer to reduce the voltage followed by a rectifier and a voltage regulator are called *linear* power supplies. An alternative design is the *switching* power supply. Instead of using a transformer to reduce the voltage, these supplies rectify the mains voltage to generate a high DC voltage. This voltage is then switched on and off at a high frequency using a fast solid-state switch, and the resulting waveform is fed through a lowpass filter to filter out the AC switching components.



Simplified Circuit Diagram of a Switching Power Supply

The 240 V AC mains supply is rectified by the full-wave bridge rectifier consisting of diodes D_1 to D_4 and smoothed by C_1 to generate 338 V DC. This DC is switched on and off at a frequency of 100 kHz or so by a high-speed electronic switch, which is shown in the circuit diagram as a simple switch. The resulting high frequency AC waveform is fed into the primary of the isolating 1:1 transformer. The purpose of this transformer is to prevent the DC output from being connected to the mains supply, rather than to perform any voltage conversion. The voltage on the secondary of the isolating transformer is half-wave rectified by D_5 and smoothed by C_2 to give a DC output.

The output voltage of the power supply depends on how much time the switch spends in the "on" position compared to how much time it spends in the "off" position. If the switch spends only a small percentage of its time in the "on" position, only a little power will be transferred to the transformer, and the output voltage will be small. If it spends a lot of time in the "on" position, a lot of power will be transferred and the output voltage will be high. In actual practice, the amount of time that the switch spends in the "on" position is controlled by electronics (not shown in the diagram) that continually monitors the output voltage and adjusts the duty cycle of the switch in order to maintain a constant output at the desired voltage. This control system uses *feedback*, where the output is monitored to control how the input is handled.

A switch-mode power supply is pretty complex compared to a simple linear supply, so there must be some significant advantages to make it worthwhile. The main advantages of the switching supply are that

1. It dissipates very little power. The switch is either "on", in which case there is current flowing through it but little voltage across it; or "off" in which case there is a high voltage across it but no current flowing through it. In either case the power dissipation is minimal compared to the linear voltage regulator, which has a voltage drop across it and a current flowing through it at the same time, and so is continuously dissipating power.

2. Because the transformer in a switching supply operates at a much higher frequency than mains-driven linear supplies—usually around 100 kHz instead of the 50 or 60 Hz of a standard mains supply—it can be physically very small and light. The size of the core required in a transformer decreases as the frequency increases. The

output filter circuit can also be much smaller and lighter, as the capacitor discharge cycle is very small and much less capacitance is required than for a 50 Hz supply.

As a result, switching power supplies are generally smaller, lighter and more efficient than their linear counterparts, and can often run off any mains voltage without having to change a selector switch. However, they also have their disadvantages. In particularly, poorly designed switching supplies can generate a lot of RF interference, which is a real problem for amateur radio applications. However, well designed and properly shielded switching supplies do not necessarily cause interference. Because of their high power requirements and space limitations, virtually all personal computers use switching power supplies. Some of these supplies can be adapted as general purpose power supplies for amateur radio use.

Switching supplies are very difficult to design and build and can be quite dangerous due to the high voltages they use. Although the linear power supplies in earlier sections make good projects for amateurs, the design and construction of switching power supplies should be left to professionals!

16.4 Batteries

Many items of radio equipment are powered by batteries, so radio amateur must know a few things about them.

Most batteries are composed of several cells. Cells have an individual voltage determined by the nature of the chemical reaction inside. These cells are connected in series or parallel to make up a battery.

Individual cells normally have a voltage of between 1 and 2 V. Lead-acid cells, as used in car batteries, have a voltage of about 2 V. NiCd cells, as used for many small consumer gadgets, have a cell voltage of 1,2 V. The nominal voltage of the battery determines how many cells are required in series. A 12 V car battery has six cells, while a 12 V battery will require 10 NiCd cells.

The capacity of a battery is specified in terms of the amount of charge it can hold. Although the SI unit for charge is coulomb, batteries are specified in ampére hours (A.h). If a battery can deliver 10 A for two hours, it has a capacity of 10 A.h. Smaller batteries are normally specified in mA.h.

Better capacity or lower internal resistance can be achieved by connecting cells in parallel. Cells with the same voltage must be used, otherwise the higher-voltage cell will run flat while trying to bring the other cell up to its own voltage.

Recharging

Rechargeable batteries must be charged at a somewhat higher voltage than their nominal operating voltage. 12 V batteries are normally charged at around 14 V. The nominal voltage for 12 V automotive electrical systems is 13,8 V, which is the voltage that most mobile transceivers are designed for.

The recharge rate should be carefully controlled. Charging at too high a rate results in damage to the battery, as its internal structure is heated and possibly deformed. Too low a rate means that the charge will take too long, possibly never reaching full capacity. A simple guideline is not to exceed the current that would charge the battery in four hours. For a 10 A.h battery, the maximum charge rate would accordingly be 2½ A.

Because batteries are not 100% efficient, some of the energy stored in the battery during charging goes to waste. A 10 A.h battery will require more than 10 A.h to charge to capacity; perhaps 20 A.h may be required.

Modern intelligent chargers are highly recommended. They monitor the actual response of the battery to the charging process, and adjust the voltage and current accordingly. In this way, they prevent damage to the battery and allow it to achieve its optimal lifetime and efficiency. The optimal current profile is very dependent on the battery technology being used. Once the battery is fully charged, these chargers will maintain a small charge current, called a trickle charge, to keep the battery topped up.

Do not attempt to recharge unrechargeable batteries. They may heat up or explode.

Summary

Linear power supplies use transformers to reduce the mains voltage, rectifiers to convert the AC to DC and an output filter including a smoothing capacitor to reduce the ripple to acceptable levels. In power supplies that use a half-wave rectifier, the ripple is at the same frequency as the mains, while in power supplies that use full-wave rectifiers the ripple is at twice the mains frequency. In South Africa, the mains supply is at 50 Hz.

A voltage regulator serves two purposes: to keep the output voltage constant despite fluctuations in the input voltage or load current; and to further reduce ripple. Integrated circuit voltage regulators may also limit current and power dissipation by the regulator to safe levels.

Switching power supplies work by rectifying the mains supply and then switching it on and off at a high frequency. The voltage output is regulated by changing the duty cycle of the switching waveform; that is, the percentage of the time that the switch is “on”. Although most switching supplies still use transformers to isolate the output from the mains supply, these transformers can be small and light because they operate at high frequency, typically around 100 kHz. The advantages of switching supplies are that they are smaller, lighter and more efficient than linear supplies. However, poorly designed switching supplies can generate a significant amount of RF interference.

Storage cells can be combined to form a battery. Cells in series provide a higher voltage. Cells in parallel can provide more current. Capacity is measured in A.h or mA.h. Charging requires careful monitoring to achieve safety, efficiency and long lifetime.

Revision Questions

1 The ripple frequency appearing at the output of an AC-fed power supply using a full wave rectifier will be:

a. Twice the input frequency.
b. Half the input frequency.
c. The same as the input frequency.
d. Dependent on the number of rectifier diodes.

2 To obtain a full-wave rectified output from a transformer using two diodes the transformer must be:

a. An isolation transformer.
b. A step-down transformer.
c. Centre tapped on the secondary winding.
d. Earthed.

3 By introducing a smoothing capacitor and an inductor in a power supply output:

a. The output voltage will increase.
b. The load can be increased.
c. The output voltage will be regulated.
d. The ripple will be reduced.

4 A smoothing circuit using an inductor and capacitor is a standard:

a. Lowpass filter.
b. Voltage regulator.
c. Rectifier.
d. Discriminator.

5 A voltage regulator in a power supply can provide:

a. Reduced ripple at the output.
b. Short-circuit protection.
c. Stable output voltage.
d. All of the above

6 A zener diode is used in a power supply to:

a. Stabilise a reference voltage.
b. Load an output circuit.
c. Introduce a noise signal.
d. Prevent excessive current flow.

7 Smart battery chargers are recommended because:

a. They allow non-rechargeable cells to be recharged.
b. They charge more rapidly.
c. They preserve the battery by not exceeding its recommended charge rate.
d. They look great in a tuxedo.

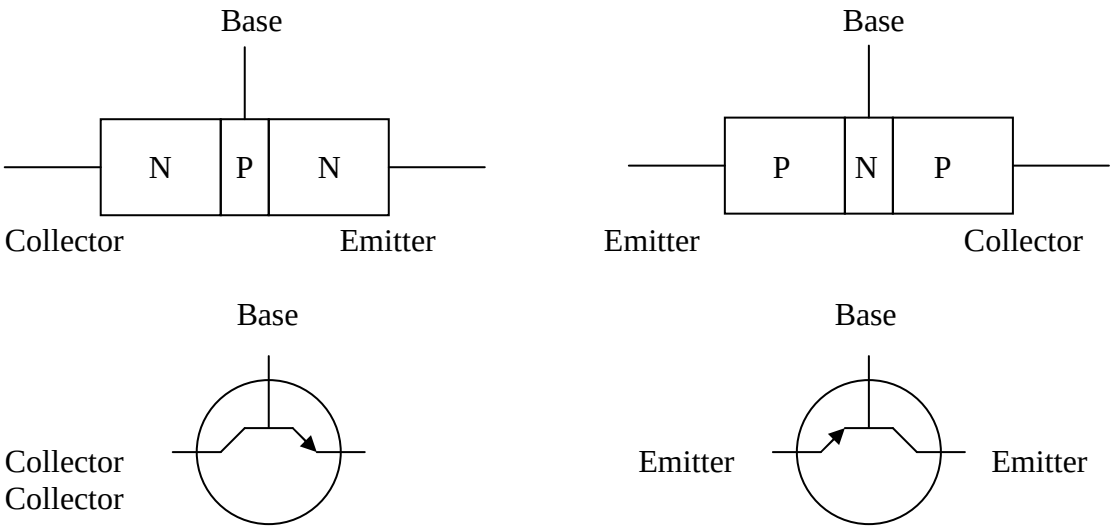
8 A battery of a specific voltage can be made by:

a. Using cells with the correct characteristic voltage.
b. Using multiple cells in series.
c. Using multiple cells in parallel.
d. Charging a cell to the correct voltage.

Chapter 17: The Bipolar Junction Transistor

17.1 Types of Transistors

A bipolar junction transistor ("transistor" for short) consists of a thin layer of P-type or N-type semiconductors called the "base", which is sandwiched between two thicker layers of semiconductor, the "collector" and "emitter" with the opposite polarity. Transistors come in two polarities: NPN transistors, which have a P-type base sandwiched between an N-type emitter and collector; and PNP transistors, which have an N-type base sandwiched between a P-type emitter and collector. The construction and circuit symbols of NPN and PNP transistors are schematically shown below:



The Physical Structure and Circuit Symbol for NPN and PNP Transistors

Note that the terminal with the arrowhead in the circuit symbol is always the emitter. The arrow represents the base/emitter junction, which has similar properties to a diode, and shows which direction the Base/Emitter and Collector/Emitter current flows in.

17.2 Operation of the NPN Transistor

To understand the operation, assume that an NPN transistor like the one on the left has a potential difference applied between the collector and emitter, making the collector positive with respect to the emitter. Then at the junction between the collector and the base, an N-type semiconductor meets a P-type semiconductor, creating a reverse-biased diode junction. Just as in a diode, free electrons from the N-type material will migrate across the junction, filling the holes in the P-type material, and creating a thin depletion layer that will prevent current from flowing.

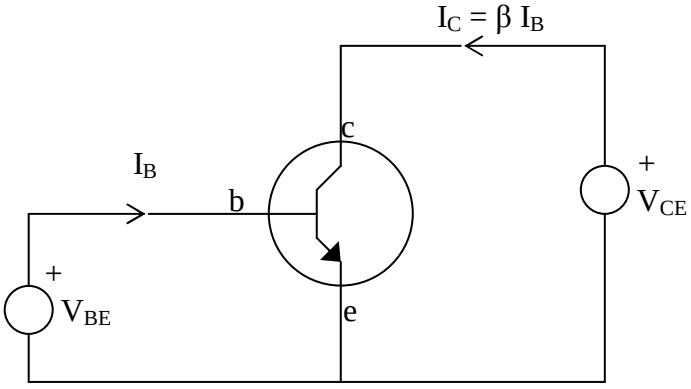
Now suppose a potential difference is applied between the base and the emitter, making the base more positive than the emitter. The PN junction acts like a forward-biased diode, so provided the base-emitter voltage exceeds to forward bias voltage for this junction (about 600 mV for silicon transistors) a current can flow from the base to the emitter. This current is carried by electrons from the emitter that are attracted by the positive potential of the base and cross the junction into the base where they combine with holes.

However, because the base is very thin (much thinner than shown), many of the electrons from the emitter that are attracted by the base potential do not end up colliding with holes and recombining, but instead make it all the way across the base and into the collector. Since electrons are now moving from the emitter to the collector, there is a current flow

from the collector to the emitter despite the reverse-biased collector/base junction! This is possible because electrons from the emitter that escape recombining with holes in the base act as charge carriers in the depletion layer of the reverse-biased junction, allowing a current to flow.

So in an NPN transistor, making the base 600 mV or so more positive than the emitter will allow a current to flow both from the base to the emitter and from the collector to the emitter, provided of course that the collector is also positive with respect to the emitter. It turns out that transistors can be designed so that the current that flows from the collector to the emitter (known as the "collector" current) is many times greater than the current from the base to the emitter (the "base" current). This enables transistors to make small signals larger, a process called amplification.

The ratio of collector current to base current is known as the beta or "current gain" of the transistor, and is represented by the Greek lower-case letter beta (β). In ordinary small signal transistors (designed for low-power work), β typically ranges from 100 to 500. One limitation is that β is not well controlled, so two transistors from the same batch may have quite different values of β. For this reason it is important to design circuits that do not rely on a particular value of β for correct operation.



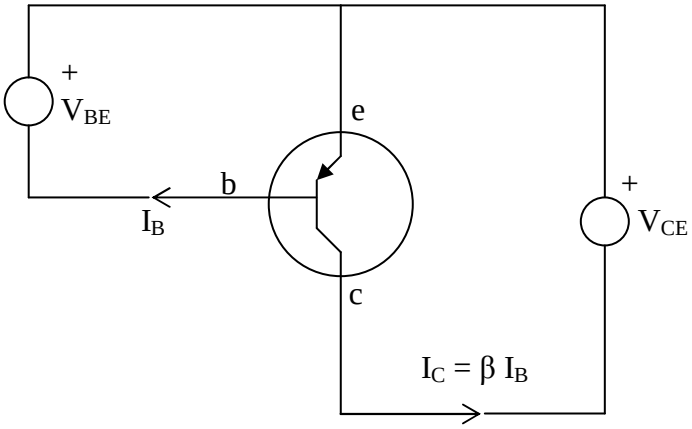
Operation of the NPN Transistor

The operation of the NPN transistor can be summarised as follows:

- The collector should always be kept positive with respect to the emitter.
- If the base/emitter voltage V_{BE} is less than 600 mV, no base or collector current flows, and the transistor is "shut off".
- Once V_{BE} reaches 600 mV, a base current I_B flows, causing a larger collector current $I_C = \beta I_B$ to flow.
- V_{BE} will remain around 600 mV as long as any base current is flowing.
- The value of β ranges between 100 and 500 for typical small-signal transistors, but is not well controlled and should not be assumed to have a particular value.

17.3 Operation of the PNP Transistor

The PNP transistor operates similarly to the NPN transistor, but with the opposite polarity. When the base gets 600 mV more *negative* than the emitter, a base current I_B and a larger collector current $I_C = \beta I_B$ flow.



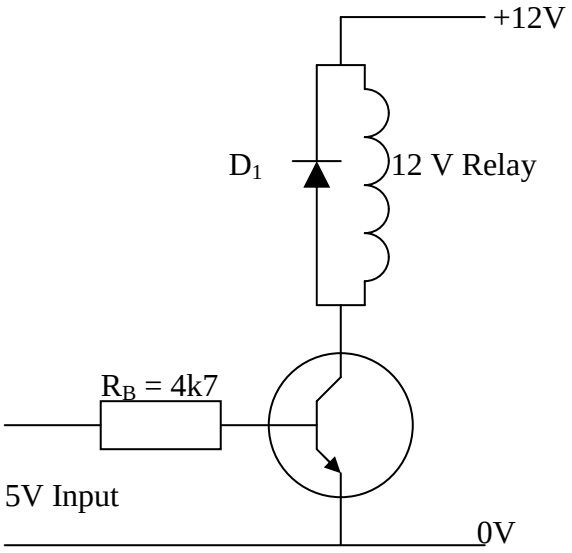
Operation of the PNP transistor

The operation of the PNP transistor can be summarised as follows:

- The collector should always be kept negative with respect to the emitter.
- If the base/emitter voltage V_{BE} is greater than 600 mV (with negative polarity), no base or collector current will flow, and the transistor is "shut off".
- Once V_{BE} reaches -600 mV, a base current I_B flows, causing a larger collector current $I_C = \beta I_B$ to flow.
- V_{BE} will remain around -600 mV as long as any base current is flowing.
- The value of β ranges between 100 and 500 for typical small-signal transistors, but is not well controlled and should not be assumed to have a particular value.

17.4 The Transistor Switch

Transistors are often used as switches because a small base current can turn the transistor "on" and allow a large collector current to flow. For example, suppose you want to switch a 12 V relay that draws 20 mA from a 5 V microprocessor control signal that can only supply 1 mA. You could use the following circuit:



A transistor switch driving a relay coil

When the input signal is off (0 V), no base current flows so the transistor is turned off and no collector current flows. When the input signal is turned on (+5V), it raises V_{BE} to

600 mV so the voltage across the base resistor R_B is 5 V – 600 mV = 4,4 V, so the current flowing through R_B and into the base I_B = 4,4V/4700 Ω = 940 °µA. This current is sufficient to turn the transistor "on", causing a collector current to flow and operate the 12 V relay. The Diode D_1 prevents the back EMF from the inductance of the relay coil from destroying the transistor when the relay is turned off again.

Transistor switches like this are usually used to switch DC voltages and currents. When high frequency signals need to be switched, diodes or relays are usually used. Of course the relay driver circuit might be similar to the one above.

Summary

Bipolar Junction Transistors are semiconductor devices that consist of a thin *base* made of either P- or N-type material sandwiched between a *collector* and *emitter* made of the opposite polarity semiconductor.

In an *NPN* transistor, both the collector and base are made positive with respect to the emitter. If the base/emitter voltage is less than 600 mV, the transistor is turned off and no collector current flows. When the base/emitter voltage reaches about 600 mV, a small base current and a large collector current will flow.

In a *PNP* transistor, both the collector and base are made negative with respect to the emitter. If the base/emitter voltage difference is greater than 600 mV, the transistor is turned off and no collector current flows. When the base/emitter voltage reaches about -600 mV, a small base current and a large collector current will flow.

The ratio of the collector current to the base current is called the "beta" of the transistor, and typical values for small-signal transistors range between 100 and 500. As long as any base current is flowing, the base/emitter voltage will remain around 600 mV for NPN silicon transistors, or –600 mV for PNP silicon transistors, irrespective of the actual base or collector current. Since transistor betas may vary widely, even for transistors of the same type, circuits should not rely on a specific value of beta.

Transistors can be used as DC switches, to allow a low voltage or small current (or both) to switch a larger voltage and current.

Revision Questions

1 If the base potential of a NPN transistor is held at the emitter potential, the collector current will be

a. Zero.
b. Always 1 Amp.
c. Between 10 mA and 2A.
d. Very high.

2 For a silicon transistor to conduct:

a. The base-emitter must be forward-biased by 0,6V.
b. The base must be connected to the emitter.
c. The collector must be connected to the emitter.
d. The base lead must be disconnected.

3 The beta of a transistor is

a. The ratio of the collector current to the base current.
b. The ratio of the collector voltage to the base voltage.
c. The ratio of the collector current to the emitter current.
d. The ratio of the collector voltage to the emitter voltage.

4 For a collector current to flow in a PNP transistor

a. Both collector and base must be positive with respect to the emitter.
b. Both collector and base must be negative with respect to the emitter.
c. The collector must be positive and the base negative with respect to the emitter.
d. The collector must be negative and the base positive with respect to the emitter.

5 If a transistor is used to control a relay that does not have protection diodes, the back EMF from the relay solenoid may

a. Increase the switching time
b. Decrease the switching time.
c. Deduce the power consumption.
d. Damage the transistor

6 If the base current in a transistor is 100 µA and the beta of the transistor is 100, the collector current is

a. 1 mA
b. 10 mA
c. 100 mA
d. 1 A

7 For an NPN transistor in normal operation

a. The collector voltage exceeds the emitter voltage
b. The emitter voltage exceeds the collector voltage
c. The emitter voltage exceeds the base voltage
d. The collector and emitter voltages are equal.

Chapter 18: The Transistor Amplifier

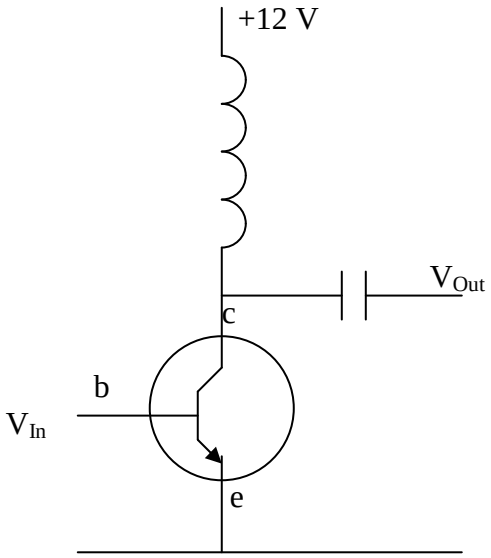
18.1 Amplification

Amplification is the process of increasing the power of a signal. Since power is V x I, this process will involve increasing either the voltage, or the current, of the signal, or possibly (but not necessarily) both. It is quite possible to amplify a signal without increasing the voltage of the signal, provided that the current is increased. Conversely, it is possible to increase the voltage of a signal without amplifying it. In the case of the step-up transformer, although the voltage of the signal is increased, the current is decreased so the power remains the same and no amplification takes place.

Amplification is crucial for radio receivers. The radio signals received from the antenna may be as small as –130 dBm (10^{-16} W). In order to make them audible, the receiver must convert them into signals in the order of 0 dBm (1 mW), which requires amplification by a factor or 10^{13}. In actual fact the amplification of a radio receiver is often greater than this[3], to make up for losses in other components like filters.

18.2 Class C Amplifiers

Because the transistor allows a large collector current to be controlled by a small base current, transistors are used in many amplifiers. Let us consider a simple design for a transistor amplifier to work at radio frequencies.



Simple transistor amplifier

Here, the input signal is applied directly to the base of the transistor and the output signal is taken from the collector. This circuit is called a “common emitter” amplifier, because the emitter of the transistor is common to both the input and output circuits.

The inductor and capacitor play two roles. Firstly, the inductor allows DC to pass while blocking radio frequency (RF) signals. The collector of the transistor can now be *biased* so that it is positive with respect to the emitter, as required for correct operation. It also provides the route for supply current to flow from the +12 V source to the amplifier—since we are making the signal more powerful, and power cannot be manufactured from thin air, it must come from the +12V supply connected via the inductor to the collector. The output

[3] 10^{13} = 10 000 000 000 000, or a more than a thousand times the human population!

capacitor allows the amplified AC signal to pass, while blocking the DC supply voltage, preventing the bias voltage from interfering with the stage that follows this amplifier.

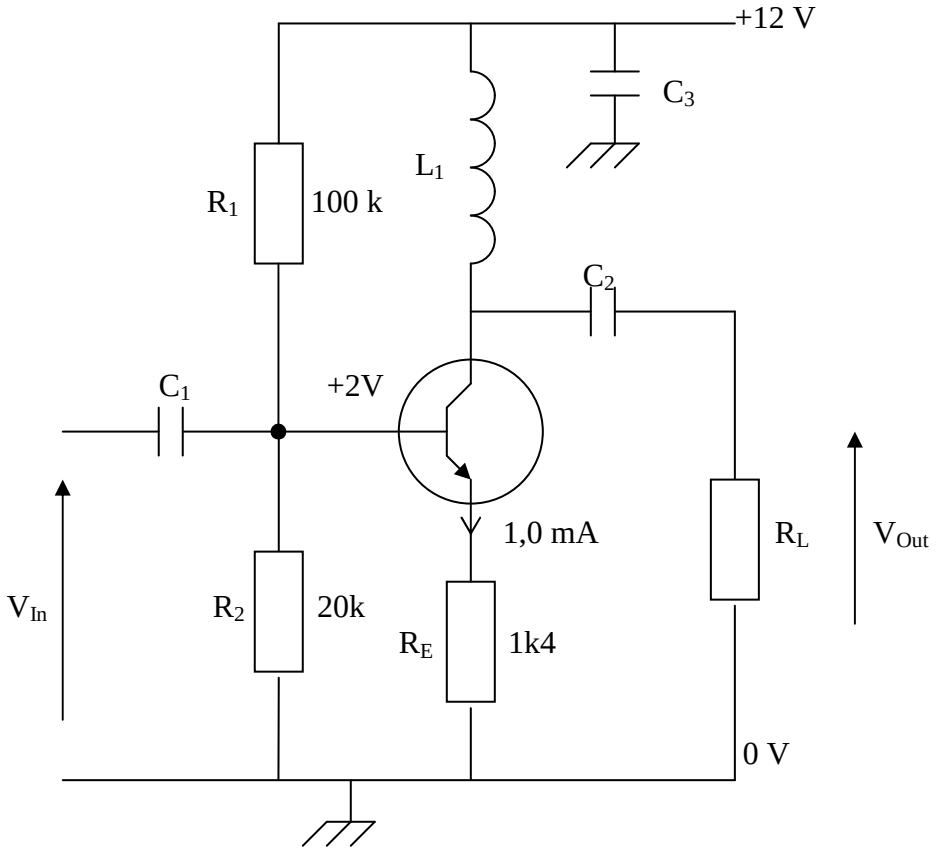
Let us consider the operation of this circuit. If the input signal is less than 600 mV, the base-emitter voltage for a silicon transistor, no current will flow into the collector, and the collector voltage will be +12 V. If the input signal exceeds 600 mV, the transistor will start to conduct, with the collector current equal to the transistor ß times the base current. The inductor will oppose any sudden change in the flow of current by generating a voltage across itself that will reduce the collector voltage. When the input signal drops below 600 mV, the transistor stops conducting. By this time there will be some current flowing through the inductor, and the inductor will oppose any sudden change to the flow of current by generating a voltage across itself in the other direction, this time raising the collector voltage above the 12 V supply voltage, possibly to as much as 24 V (twice the supply voltage). When the input signal again exceeds 600 mV and the transistor starts to conduct, the collector voltage will again drop, and so on. These variations in the collector voltage constitute an AC signal that will be passed through the capacitor as an output signal.

In this amplifier, the transistor does not conduct the whole time. In fact, it conducts somewhat less than half of each AC cycle, since it will only start to conduct when the input voltage exceeds 600 mV. An amplifier that conducts for less than one half of a cycle is called a “Class C” amplifier. Since class C amplifiers do not reproduce the shape of the input waveform accurately, they generate a large amount of distortion. However, they do have the advantage of being quite efficient compared to other amplifiers, since most of the power supplied to the amplifier (in this case from the +12 V supply) ends up in the output signal. Efficiencies of 60 to 70% are common for Class C amplifiers. Also note that there is not much point trying to amplify a signal with a peak voltage of less than 600 mV using this amplifier, since it won’t ever be sufficient to start the transistor conducting!

So what would one do with an amplifier that requires a large input signal and generates considerable distortion, but is relatively efficient? Class C amplifiers are often used as the RF power amplifier for CW (Morse code) and FM transmitters, since in these applications it turns out that the distortion (also called *nonlinearity*) of the amplifier is not a big problem as the unwanted harmonics can be filtered off quite easily by a lowpass filter at the output. However, Class C amplifiers are *not* suitable for use in single sideband (SSB) or amplitude modulation (AM) transmitters; for these you need a *linear* amplifier. Don’t worry if you are not familiar with the terms CW, AM, FM and SSB, they will be explained in a later module.

18.3 The Class A Common-Emitter Amplifier

In order to faithfully amplify small signals, we usually use Class A amplifiers. In this class of amplifier, collector current flows in the transistor throughout the entire cycle of the input waveform. A typical circuit might be as follows:



Class A common-emitter amplifier

Here, R_1 and R_2 form a *voltage divider* that applies a certain *bias voltage* to the base. For example, suppose that R_1 is 100 kΩ and R_2 is 20 kΩ, the bias voltage at the base will be 2 V. Since this voltage is greater than 600 mV, it is sufficient to cause a current to flow from the base to the emitter, and we know that the base/emitter voltage will be about 600 mV. The voltage across R_E, the emitter resistor, is then 2 V – 600 mV = 1,4 V. Suppose R_E is 1,4 kΩ, then the current flowing through R_E (and also through the emitter of the transistor) is 1,4 V ÷ 1,4 kΩ = 1,0 mA.

So how much of this total current is base/emitter current, and how much is collector/emitter current? The answer depends on the ß of the transistor. If ß= 99, the collector/emitter current is 99 times the base/emitter current, so 1% of the emitter current comes from the base, and 99% from the emitter. In this example, the collector current would be 990 µA.

But hold on a moment. We have already been warned that the ß of a transistor is not well controlled and should not be relied upon to have a specific value. What if ß = 499 instead of 99? The collector current will then be 499/500 of the emitter current, or 998 µA, instead of 998 µA. As you can see, the collector current is constant to within 1%, despite variations in the ß from 99 to 499. For most practical purposes, the collector current can be assumed to be equal to the emitter current, with the base current being ignored (although of course the base current is very important in practice, as it is what allows the collector current to flow!)

Finally, let us consider the input and output impedance of the amplifier. An RF input signal will "see" an input impedance consisting of resistors R_1 and R_2 in parallel. Why in parallel?

The supply voltage (in this case +12 V) always has an AC path to ground. In this case, it is provided by the "decoupling" capacitor C_3, which is connected between the supply voltage and the chassis. So the input impedance is 100 kΩ in parallel with 20 kΩ, or about 16,7 kΩ. There is an additional component, consisting of the emitter resistor R_E multiplied by the ß of the transistor that is also in parallel with the input, but this is usually significantly higher than the bias resistors R_1 and R_2 and can be neglected.

Having determined the input impedance, what then is the output impedance? Since the collector current, being ß times the base current, does not depend much on the collector voltage, the collector acts like a current source and its impedance is fairly high. The inductor L_1 will also be chosen to exhibit high reactance at the design frequency. So we can say that in this case the output impedance is "quite high" without actually putting a figure to it.

And what is the gain of the amplifier? Since the base/emitter voltage remains constant at 600 mV, if we change the base voltage by a small amount, which we shall call V_{IN}, the emitter voltage must also change by the same amount. The emitter current I_E must then also change by a small amount, $\Delta I_E = V_{IN}/R_E$. But since the collector current is virtually identical to the emitter current (we are neglecting the small effect of the base current), the output current will change by the same amount as the emitter current, so $I_{OUT} = V_{IN}/R_E$. The effect that this has on the output voltage (and hence the gain of the amplifier), depends on the resistance of the load, R_L. To be precise,

$$\begin{aligned} V_{Out} &= I_{Out}\, R_L \\ &= (V_{In}/R_E)\, R_L \\ &= V_{In}\, R_L/R_E \end{aligned}$$

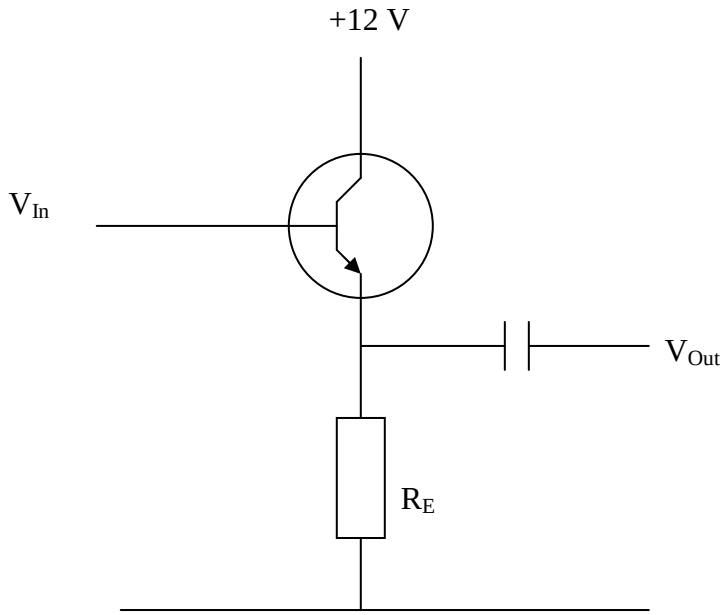
So the *voltage gain* of this amplifier is R_L/R_E. To calculate the *power gain*, need a specific value for the load resistance. Let's choose it as 16,7 kΩ, the same value as the input resistance of the amplifier. The power gain is then just the square of the voltage gain. The voltage gain will be 16,7 kΩ ÷ 16,7 kΩ = 11,9 and the power gain will be $11{,}9^2 = 142$ or 21,5 dB. This gain is typical of the gains achievable by small-signal common-emitter amplifiers.

Note that in this design, because the base of the transistor is biased to a voltage of +2 V, the transistor will conduct provided the input voltage does not go below -1,4 V. So as long as the input voltage is less than 1,4 V *peak* (about 1,0 V RMS), the transistor will conduct throughout the full cycle of the input waveform, making this amplifier Class A. It will faithfully reproduce the exact shape of the input signal at the output, and so is a *linear* amplifier that can be used to amplify SSB and AM signals without distortion.

Of course, the DC bias current flowing through the collector circuit of the transistor the whole time does waste quite a lot of power, making Class A amplifiers much less efficient than their Class C counterparts. Class A amplifiers are typically only about 25% efficient – that is, only 25% of the power supplied by the power supply actually ends up in the load. The other 75% ends up heating the output transistor!

18.4 The Common-Collector (Emitter Follower) Amplifier

Consider the following circuit:



Common-collector amplifier

Suppose that the input voltage is always between 600 mV and 12 V, so the transistor always conducts. It is therefore operating in Class A. This time, the output voltage is taken from the emitter instead of the collector of the transistor. What do we know about the output voltage? Since the base-emitter voltage is always 600 mV if the transistor is conducting, the emitter voltage must "follow" the base voltage, although it will always be 600 mV less than the base voltage. So any change in the input voltage will result in an equal size change in the output voltage, and the amplifier has a voltage gain of 1. Because the DC bias is removed by the DC blocking capacitor at the output, the output voltage is *not* necessarily 600 mV below the input voltage.

This circuit is called a "common collector" amplifier, and is also known as an "emitter follower" because the emitter voltage "follows" the base voltage.

So why would anyone want an amplifier if the output voltage is the same as the input voltage? The secret is in the impedances. The output impedance of the emitter follower is low, making it a good voltage source, because the output voltage does not depend on the load resistance. However, the input impedance is high, so it does not load the preceding stage much. It is a good circuit to use as a buffer stage, to prevent changes to the input impedance of the following stages from affecting the preceding stages. And since the low-impedance output is capable of supplying much more power than the high-impedance input consumes, the common collector amplifier is quite capable of providing a *power* gain even though it has unity *voltage* gain.

18.5 The Common Base Amplifier

There is a third transistor amplifier configuration, known as the "common base" configuration, which is less common than the common emitter or common collector configurations. It has low input impedance (typically only a few ohms) and high output impedance (making is a good current source), and a current gain of unity (one). You could consider it to be a "current follower" since the output current is identical to the input current. It can also provide a power gain if the load resistance is greater than the input impedance of the amplifier.

18.6 The Class AB Amplifier

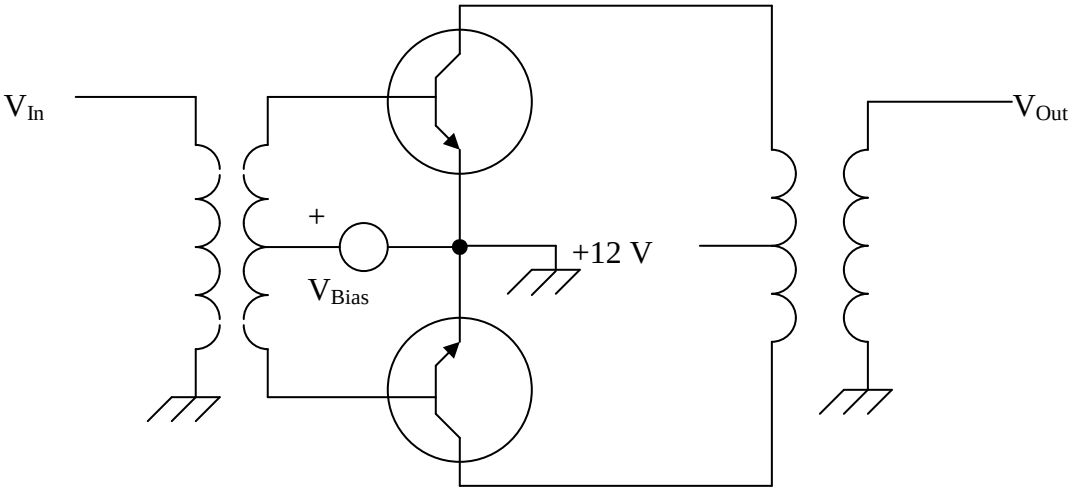
In a Class A amplifier, the transistor (or other output device) conducts for the full cycle of the input waveform. This is also referred to as conduction over 360˚ of the input cycle, a reference to the typical sinusoidal characteristics of radio signals. Class A amplifiers are linear – that is, they accurately reproduce the shape of the input waveform at the output and so introduce little distortion – but typically inefficient.

In a Class C amplifier, the transistor conducts for less than half the cycle, in other words for less than 180˚ of a sine. Class C amplifiers can be pretty efficient, but they are very non-linear, introducing a substantial amount of distortion into the output.

There is also a Class "B", where the transistor conducts for exactly half the input form (180˚). However, this type is not commonly used.

Class AB amplifiers use two output devices operating in a "push-pull" configuration. One device conducts during the positive half cycle of the input waveform, and the other device conducts during the negative half cycle. Both devices are operated Class B, meaning that they conduct for half of the input waveform (180˚). However, between them the output devices can replicate both the positive and the negative half cycles of the input waveform, as can a Class A amplifier, which is where the name "Class AB" comes from – two Class B devices operating together to provide the same effect as a single Class A device.

A simplified push-pull Class AB amplifier circuit is shown below:



Class AB push-pull amplifier

In this circuit, the bias voltage V_{Bias} is just sufficient to keep both transistors just conducting a small current when there is no input voltage. The transformer at the input acts as a "phase splitter" to convert the input signal into two signals that are 180˚ out of phase, which are applied to the bases of the transistors. When a positive signal is applied to the base of one transistor is conducts more strongly, while the voltage at the base of the other transistor is reduced so it stops conducting. The transistors switch around every half cycle, so each of the transistors is conducting for half the cycle, or about 180˚. The transformer at the output recombines the output of the two transistors, so although each is only conducting for half the cycle, the combined output of both transistors can faithfully replicate both half cycles of the input signal.

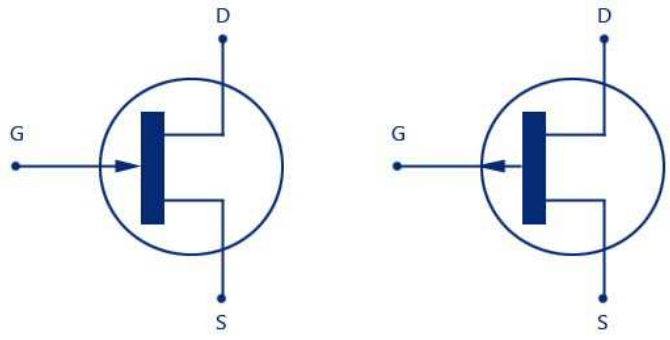
Class AB amplifiers have the advantage that because there isn't a large constant DC bias current flowing through the output devices (as there is in Class A) they are much more

efficient than Class A amplifiers; while because they reproduce both the positive and the negative half cycles of the input, they are much more linear than Class C amplifiers. Although they still introduce some distortion, such as *crossover* distortion at the point where one of the transistors stops conducting and the other starts conducting, this distortion can be kept within reasonable limits, so properly designed Class AB amplifiers are quite suitable for use as power amplifiers for AM and SSB signals.

18.7 Field-Effect Transistors

A Field-Effect Transistor (FET) is a semiconductor device that is controlled by voltage rather than by current. It is made like junction transistors, using doped semiconductor, but uses an insulating material such as aluminium oxide to provide ultra-thin insulation between different structures. This technology is known as metal-oxide semiconductor (MOS). The manufacturing process is more efficient and cheaper for large-scale integration, making FETs the prime choice for logic and computer chips.

The exact function of FETs is beyond the scope of this text. Suffice to say that there are three electrodes: Gate, Source and Drain. The application of a voltage to the gate changes the impedance between the drain and source, much like the base current controls the impedance between the collector and emitter in junction transistors. There are basically two types, depending on the type of semiconductor from which the drain-source channel is made: P-channel and N-channel.



JFET-N-Channel and P-channel Schematic Symbol

MOSFETs have become ubiquitous in high-power switching applications, almost completely displacing junction transistors.

All the transistor circuits shown in this text can be implemented with FETs. In general, the gate would replace the base, the drain would replace the collector and the source would replace the emitter. Also, the gate requires voltage bias rather than current bias, so some of the bias networks will have to be redesigned.

Even the configurations can be duplicated. Common-gate, source-follower and drain-follower ampliers can be implemented relatively simply.

18.8 Thermionic Tubes

Don’t worry about the detail in this section. It is only included for historical reasons, and in case you want to build a high-power amplifier one day.

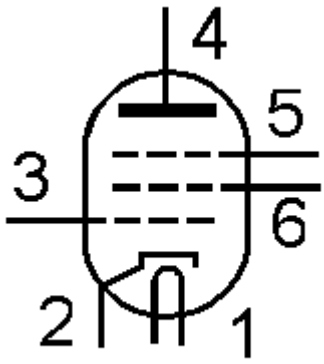
Tubes (Mericun), or Valves (British), are the oldest active devices in electronics. The thermionic diode was invented in 1873. The first radio transmissions happened around 1895. The triode was invented in 1906. Semiconductor transistors were only invented in 1947. Tubes have now largely been replaced by semiconductor devices, but survive in high-

power RF applications, such as linear amplifiers at levels of 1 kW and above. Tubes are a great vehicle for homebrew amplifiers, as they are generally more robust than most transistors.

Thermionic tubes come in different flavours. The diode consists of a heated cathode and an anode (hence the "di-!"), enclosed in a vacuum. The vacuum is normally contained in a glass or ceramic envelope. Electrons are emitted by the heated cathode, in a process known as *thermionic emission*, forming a cloud around the cathode. If the anode is positive relative to the cathode, electronics are attracted by the anode, causing a current to flow. If the anode is negative relative to the anode, no electrons jump the gap, with no current flowing.

If a grid is introduced between the cathode and the anode, a small voltage applied to the grid will control the amount of current flowing. By varying the grid voltage, the anode current can be controlled (much like a FET). Several grids can be introduced to achieve desired behaviour, including improved immunity to capacitive effects at high frequencies. The device family consists of diodes (two elements), triodes (three elements), tetrodes (four), pentodes (five) and derivatives such as the cathode ray tube (CRT) and magnetron (used to generate high-power RF in many microwave ovens).

The diagram shows a pentode. 1 is the heater. 2 is the cathode. 4 is the anode. 3, 5 and 6 are the various grids (control, suppressor and screen grids respectively).



Symbol for a pentode

Apart from high-power RF applications, tubes are popular in high-end audio equipment, including guitar amplifiers. Their harmonic distortion differs from that of solid-state devices, providing a more pleasing sound to some discerning ears.

Most of the designs in this text can be adapted from transistors to tubes, making suitable changes to the bias networks and scaling the operating voltage accordingly.

18.9 Integrated Circuits

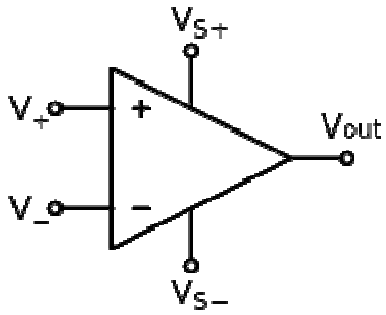
Discrete components are seldom used in modern electronics. A much more compact and cost-effective solution exists in *integrated circuits* (ICs). ICs (or "chips") pack thousands of components onto a small piece of silicon. A dual-core Itanium 2 processor includes around 1 700 000 transistors, not to mention all the associated components to provide bias and routing, on only 544 mm^2 of space (approximately 23 x 23 mm). These processors sold for around US$ 1000, or something like 17 transistors per US cent!

ICs can be very small. A sophisticated microprocessor in a 16-pin SOIC, suitable for home projects, is about 10 mm long, 6 mm wide and 2 mm high, including its connecting leads. It

draws very little current—in the region of 1 mA. It is simply not possible to make such sophisticated electronics using discrete components without spending inordinate amounts of time and money.

Typical projects are now made using a few ICs with some discrete components, or using a microprocessor that can be reconfigured to do many different things by writing some software. These processors can be bought on a small PC board which provides easy connection to the device's ports. Excellent platforms for amateur radio projects include the Arduino and Raspberry Pi product ranges.

Perhaps the most useful single building block for small-scale amplifier and buffer applications is an operational amplifier or *op-amp*. The ubiquitous LM741 op-amp is freely available at about three per US$. The amplifier offers high input impedance, high gain, low output impedance, high linearity and low power consumption. In an SOIC package, the device needs only about 5 x 6 mm of board space. The symbol is shown below:



Operational amplifier (op-amp)

The input is applied across V_+ and V_-. The difference between these two pins is amplified at V_{out}. The supply is attached to V_{S+} and V_{S-}. A supply of +5 V and -5 V can be used, or one of the pins can be connected to ground. Likewise, with the input either pin can be connected to ground. Using suitable simple networks of outside passive components (resistors, capacitors, inductors etc.), almost any imaginable signal amplifier, filter or signal generator can be built.

Summary

An amplifier is a circuit that increases the power of a signal.

The Common Emitter amplifier can have both voltage and current gain. Common emitter amplifiers have high output impedance and moderate (10 kΩ or so) input impedance.

The common collector amplifier is also known as the "emitter follower" because the output is taken from the emitter, which "follows" the voltage on the base. The common collector amplifier has unity voltage gain but can still provide power gain if the output current is greater than the input current. The common collector (emitter follower) has a high input impedance and a low output impedance.

The common base amplifier has a low input impedance and a high output impedance, and has a unity current gain (i.e. the output current is the same as the input current).

Class A amplifiers conduct for the full cycle (360˚). They have low distortion (good linearity) but are relatively inefficient. Almost all small-signal amplifiers are Class A, because efficiency is not important when so little energy is wasted.

Class B amplifiers conduct for exactly half the cycle. They are not commonly used.

Class C amplifiers conduct for less than half the cycle (less than 180˚). They can be very efficient, but are non-linear and introduce distortion. Although they can be used as power amplifiers for CW and FM signals, they cannot be used for AM or SSB signals.

Class AB amplifiers use two Class-B output devices operating in push-pull configuration to amplify both the positive and the negative half cycles. They are more efficient than Class A amplifiers, and while they also have more distortion than Class A amplifiers they can still be used as power amplifiers for AM and SSB signals.

Field Effect Transistors (FETs) are replacing junction transistors in many applications, especially in logic circuits and high-power applications, such as RF amplifiers.

Thermionic tubes have largely been replaced by semiconductors, but survive in high-power RF and some audio applications.

Integrated circuits contain a myriad of components, offering great sophistication at low prices and in very small packages. Microprocessors can be reconfigured for many different applications by writing some software. Excellent platforms for amateur radio projects include Arduino and Raspberry Pi.

An operational amplifier is a cheap, small component that can implement almost any simple analogue signal buffer, amplifier, filter or oscillator.

Revision Questions

1 The output impedance of an emitter follower buffer amplifier is:

a. Infinite.
b. Very high.
c. 0.
d. Fairly low.

2 In a transistor amplifier circuit where full base current always flows, the circuit is biased for operation in:

a. Class A.
b. Class B.
c. Class AB.
d. Class C.

3 A class C amplifier conducts over

a. The complete cycle.
b. Three quarters of the cycle.
c. Exactly half a cycle.
d. Less than half a cycle.

4 The amplifier class which has the lowest distortion figures is:

a. Class A.
b. Class B.
c. Class AB.
d. Class C.

5 An amplifier that operates under conditions of bias and supply such that conduction occurs for more than 180° but less than 360° of a complete input cycle is operating in:

a. Class A.
b. Class AB.
c. Class B.
d. Class C.

6 When an RF power amplifier is biased for a conduction angle of 360°:

a. Output current flows for only part of the input cycle.
b. Bias current never shuts off the device.
c. The average grid voltage is twice cutoff voltage.
d. RF power is produced at greatest efficiency.

7 FETs differ from junction transistors in:

a. They are controlled by voltage rather than current.
b. They are easier to manufacture.
c. They are better in high-power applications.
d. All of the above.

8 Vacuum tubes are still found in amateur radio applications because:

a. They are more practical for high-power RF amplifiers.
b. They remain part of lots of equipment in collectors' hands.
c. They are said to provide better audio quality.
d. All of the above.

9 Op-amps can be used to implement cheap and easy:

a. Small-signal amplifiers.
b. Active filters.
c. Oscillators.
d. All of the above.

Chapter 19: The Oscillator

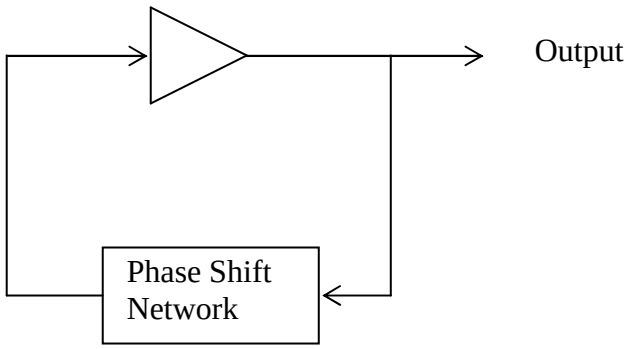
19.1 Oscillators

Oscillators are circuits that are used to generate AC signals. Although mechanical methods, like alternators, can be used to generate low frequency AC signals, such as the 50 Hz mains, electronic circuits are the most practical way of generating signals at radio frequencies.

Oscillators are widely used in both transmitters and receivers. In transmitters they are used to generate the radio frequency signal that will ultimately be applied to the antenna, causing it to transmit. In receivers, oscillators are widely used in conjunction with mixers (a circuit that will be covered in a later module) to change the frequency of the received radio signal.

19.2 Principle of Operation

The diagram below is a *block diagram* showing a typical oscillator. Block diagrams differ from the schematic or circuit diagrams that we have used so far in that they do not show every component in the circuit individually. Instead they show complete functional blocks – for example, amplifiers and filters – as just one symbol in the diagram. They are useful because they allow us to get a high level overview of how a circuit or system functions without having to show every individual component.



Block Diagram of an Oscillator

The triangular symbol at the top represents an amplifier. The input of the amplifier is the blunt side of the triangle, on the left in this diagram; the output is the pointy side of the triangle, on the right. Since this symbol always represents an amplifier, there is no need to label it. The output of the amplifier is connected to the input of the block labeled “phase shift network”, and the output of the phase shift network is connected back to the input of the amplifier. Since the rectangular box of the phase shift network does not indicate the input and output, you must surmise this from the directions of the arrows on the connecting lines. The output of the oscillator is taken from some point in the circuit. In this diagram, it comes from the output of the amplifier.

The lines connecting the symbols in the block diagram represent the flow of signals from one functional block to another. In this type of diagram, a line does not necessarily represent a single wire, as it would in a schematic (circuit) diagram. A signal might flow along a single wire (with respect to earth), or it might flow in two wires, with the current flowing in opposite directions in both wires. In either case, it could be represented by a single line in a block diagram. The arrows at the end of the lines show the direction that the signal flows in—in this case, from the output of the amplifier to the input of the phase shift network, and from the output of the phase shift network back to the input of the amplifier. The direction in which the signal is flowing does not in general correspond to the direction in which *current* is flowing. In fact, most of the signals we deal with will in any case be AC so current flows in *both* directions.

So how does this circuit oscillate? When it is initially turned on, there will be some (very small) *thermal noise* present in the circuit. This type of noise is generated by the random motion of electrons due to heat, and exists in all conductors. Thermal noise is broadband in nature, meaning that it includes frequency components at all possible frequencies. When you turn the volume of a hi-fi amp up without any input signal, the hiss you hear is the audio frequency component of the thermal noise. If you hear a hum, this is mains pickup, not thermal noise.

Thermal noise at the input to the amplifier will be amplified, causing a larger noise signal at the output of the amplifier, some of which is bled off to the output, and some of which is applied to the phase shift network. The phase shift network does what its name implies – it changes the phase of the input signal, so the output of the network will have a phase that either leads or lags the input signal. The phase relationship between the output and the input depends on the precise frequency of the input signal.

At most frequencies, the output of the phase shift network, which is fed back into the amplifier, will not be at precisely the same phase as the noise component that caused it in the first place. In this case, the signal that is "fed back" to the input of the amplifier from the phase shift network will partially cancel out the signal that caused it, so the noise components at these frequencies will die out. However, at one specific frequency, the output of the phase shift network will be exactly in phase with the noise component that caused it, and so it will reinforce that particular frequency component of the noise signal at the input to the amplifier.

This reinforced signal will again be amplified by the amplifier, phase shifted by the phase shift network, and fed back to the input of the amplifier. Once again, the output from the phase shift network is precisely in phase with the input signal from the "last round" that caused it, and so the signals reinforce each other and keep on growing.

Of course the signal cannot grow larger forever. As the signal grows bigger, ultimately the gain of the amplifier will be reduced (for example, it may be limited by the power supply voltage to the amplifier) until we reach the point that the amplified signal that is passed through the phase shift network and back to the input of the amplifier is only just as strong as the input to the amplifier that caused it. At this point, the signal is no longer growing, but remains constant and we have reached a stable oscillating state. If the oscillator has been designed correctly, the output will be a constant amplitude signal at the desired frequency.

Feeding back some of the output of the amplifier back to the input in such a way that it reinforces the original input signal is called *positive feedback*. This is the same effect that you get when the audio output of a PA system is fed back to the microphone creating "howl-around" or "feedback".

19.3 The Barkhausen Criteria for Oscillation

The *loop gain* of an oscillator is the total gain that the signal experiences starting from any point in the circuit and going around the loop until it gets back to the starting point. For example, suppose the amplifier has a gain of 10 dB, that half the power is "bled off" to the output (resulting in a loss of 3 dB), and that the phase shift network also has a loss of 3 dB. Converting the losses into negative gains, we get the following figures:

Amplifier	10 dB
Loss of output signal	-3 dB
Phase shift network	-3 dB
Total loop gain	4 dB

Similarly, you can calculate the total phase shift around the loop. The amplifier will contribute some phase shift, and the phase shift network will contribute some more. Even the interconnecting wires may contribute significant phase shift at high frequencies. The wavelength of a 100 MHz signal is 3 m, so every 1 cm of connecting wire would contribute a phase shift of about 1,2˚! At microwave frequencies, even a few mm of PC Board track can introduce significant phase delays.

When the oscillator is oscillating stably—that is, with constant amplitude and frequency—the following criteria must be fulfilled:

- The *loop gain* must be exactly 1, or 0 dB. If the gain was more than 1, the amplitude of the output would be increasing. If less than 1, the amplitude would be decreasing.

- The *total phase shift* around the loop must be 0° or an integer multiple of 360˚. This condition is necessary for the signal to reinforce itself as it goes around the loop, so it does not cancel itself out.

These requirements are known as the *Barkhausen criteria* for oscillation.

It is entirely possible for these criteria to be met at more than one frequency. In particular, it is easy for the phase requirement to be met, since it only specifies a phase shift of 0° or any integer multiple of 360˚, so it could be satisfied for different frequencies that had a total phase shift around the loop of say 0˚, 360˚ and 720˚. If both criteria are met for several frequencies, the oscillator will oscillate at all these frequencies simultaneously, which is usually not the desired result! Oscillations at undesired frequencies are called *parasitic* oscillations.

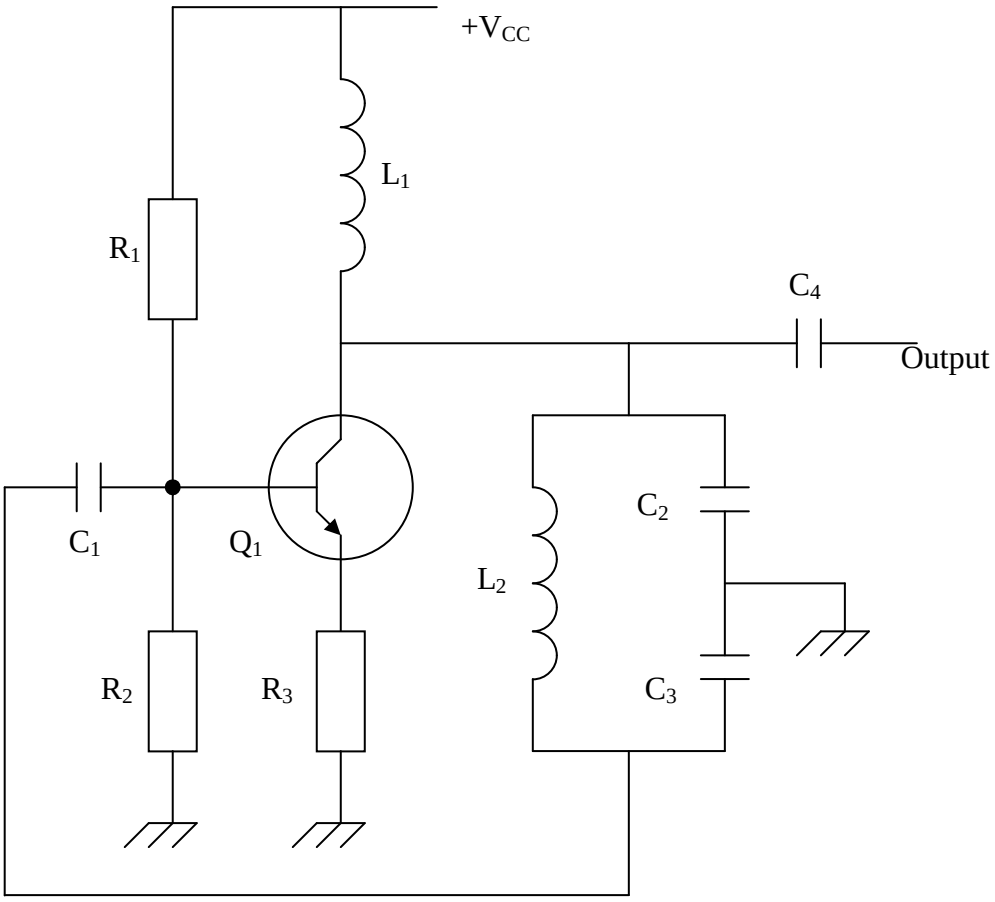
In order to minimise the likelihood of such parasitic oscillations, the phase shift network is usually also made frequency selective. It passes frequencies in the region of the desired frequency of oscillation, while attenuating higher or lower frequencies. In other words, it is made to be a *bandpass filter* as well as a phase shift network. The advantage of this arrangement is that even if the phase shift criterion is met for some other frequencies, as long as they are far enough away from the desired frequency, they can be attenuated sufficiently by the bandpass characteristic of the network to ensure that the loop gain remains less than 1 so that oscillation will not occur at these unwanted frequencies.

Fortunately, there is a simple circuit that provides both a phase shift and bandpass filter characteristics simultaneously: the parallel tuned circuit. At the resonant frequency, the reactance of a parallel tuned circuit changes rapidly from being highly inductive just below the resonant frequency to being highly capacitive just above the resonant frequency. This sudden change in reactance results in a change in the phase relationship between the voltage across the tuned circuit and the current flowing through it. Remember that for in inductive reactance, voltage leads current, while for a capacitive reactance current leads voltage. At the same time, the parallel tuned circuit can be used to provide good bandpass filter characteristics, minimizing the likelihood of parasitic oscillation.

An oscillator that uses a tuned circuit as its phase shift network will oscillate at (or very close to) the resonant frequency of the tuned circuit.

19.4 The Colpitts Oscillator

The Colpitts oscillator is typical of how these concepts can be implemented in a practical circuit.



A Colpitts oscillator

Transistor Q_1 and the associated components R_1, R_2, R_3 and L_1 form a common-emitter amplifier. The output of the amplifier, taken from the collector of Q_1, is fed into a parallel tuned circuit consisting of L_2, C_2 and C_3. The capacitor in this tuned circuit has been "split" into two capacitors, C_2 and C_3, to allow the output current from the collector of Q_1 to flow to ground via C_2. This current causes a voltage across the whole parallel tuned circuit, also known as the *tank circuit* of the oscillator. This voltage is fed back to the input of the amplifier via C_1. The output of the oscillator is taken from the collector of the transistor via C_4. The label "V_{CC}" represents the positive power supply voltage.

The defining characteristic of the Colpitts oscillator –what makes it a Colpitts oscillator as opposed to any other type of oscillator – is the way the tank circuit (the parallel tuned circuit) uses a split capacitor to allow the output of the amplifier to be injected across one of the capacitors, while the input to the amplifier is taken from across the other capacitor.

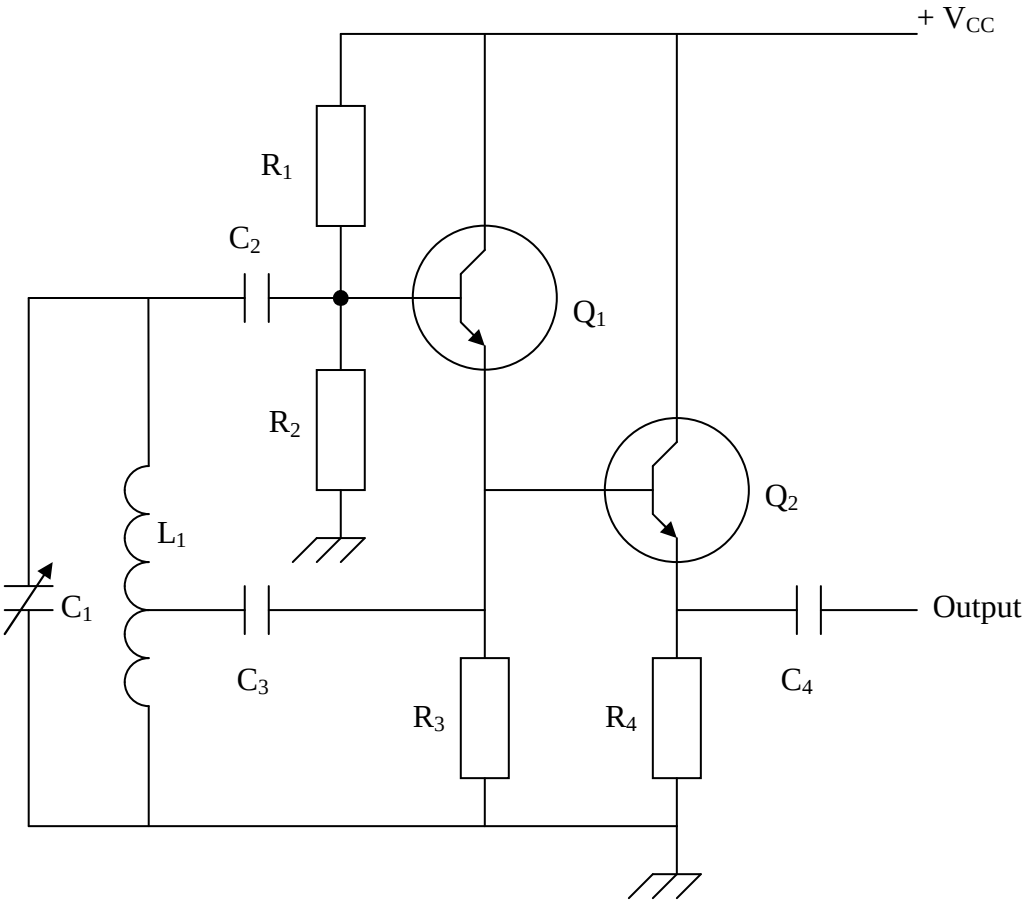
19.5 Buffering

Because the amount of signal that is drawn off by the output of the oscillator affects the loop gain of the oscillator, it will also affect the frequency of the oscillator. For this reason it is important that the amount of signal drawn off does not change, for example in response to a Morse code (CW) transmitter being keyed, otherwise the frequency of the transmitter will change as it is keyed, a phenomenon known as "chirp". Most transmitter designs prevent this by having a *buffer amplifier* between the oscillator and the keyed stages of the

transmitter. The buffer amplifier is often a common-collector (emitter follower) amplifier, which shows constant high impedance to the oscillator while having a low output impedance that can supply sufficient current to drive the stages that follow.

19.6 The Hartley Oscillator

Another way of feeding the output of the amplifier into a parallel tuned circuit, and the output of the tuned circuit back to the input of the amplifier, is to use a centre-tapped inductor in the tank (tuned) circuit. This principle defines the Hartley oscillator.



Hartley Oscillator with a buffer amplifier

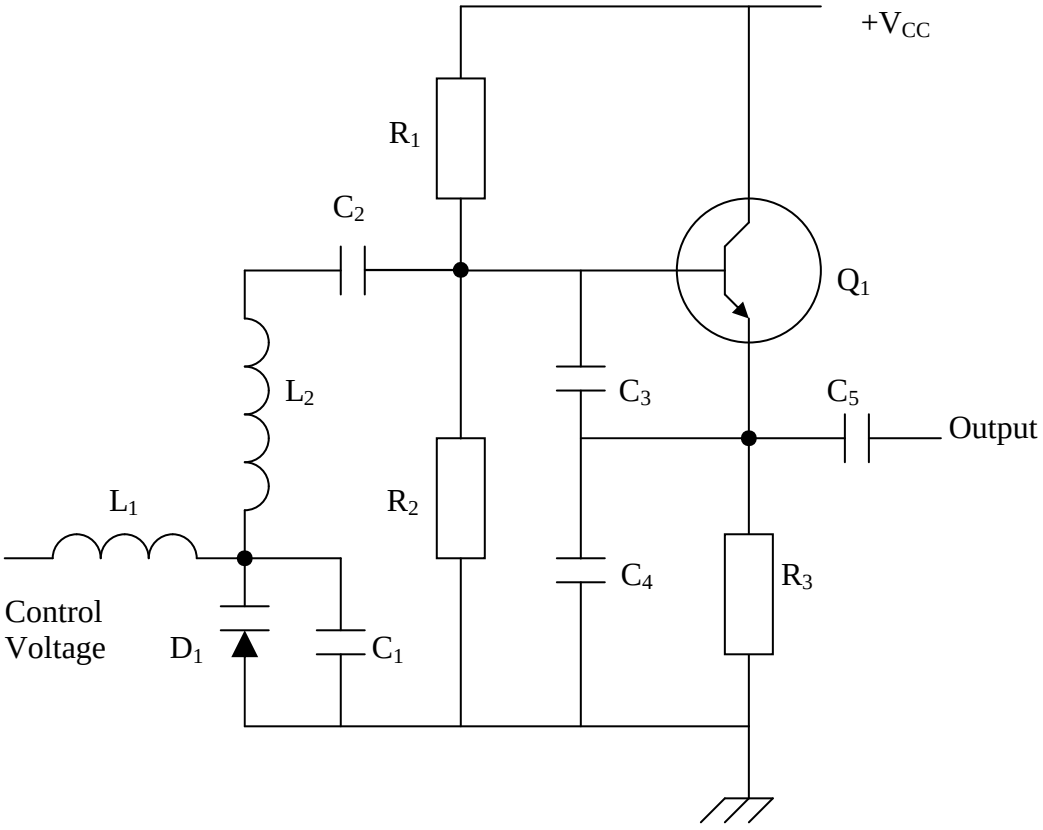
In this circuit, transistor Q_1 is a common-collector (emitter follower) amplifier that is biased by R_1, R_2 and R_3. The output of the amplifier, at the emitter of Q_1, is coupled via DC blocking capacitor C_3 into the parallel tuned tank circuit consisting of L_1 and C_2 through a tap in the inductor. The tank circuit is coupled back to the input of the amplifier via C_2, which serves as another DC blocking capacitor to prevent the base of Q_1 from being shorted to earth via L_1. The arrow through C_1 indicates that it is a variable capacitor, so the resonant frequency of the tank circuit, and hence the oscillator frequency, can be changed by varying C_1. The output of the oscillator at the emitter of Q_1 is fed to Q_2, which is a common-collector (emitter follower) buffer amplifier. R_4 sets the emitter and collector current for Q_2. The output of the buffer amplifier is taken from the emitter of Q_2 via DC blocking capacitor C_4. An oscillator where the frequency can be varied, typically by turning a control knob, is known as a *Variable Frequency Oscillator* (VFO).

In this circuit, the centre-tapped inductor L_1 acts a bit like a step-up transformer, since an AC voltage applied between the centre tap and the chassis connection (the bottom of the

inductor) generates a varying magnetic field, which causes a larger voltage to be generated between the "hot" side of L_1 (the top of the inductor) and the chassis. This voltage step-up allows the common-collector amplifier to provide power gain in this circuit, despite the fact that the voltage gain between the base and emitter of the transistor is unity (1). A tapped inductor like this is also called an *autotransformer*.

19.7 The Voltage-Controlled Oscillator

If part of the capacitance forming the tuned circuit in an oscillator is made up of capacitance from a varicap diode, the frequency of the oscillator can be varied by changing the reverse-bias voltage applied to the varicap diode. This configuration is called a *voltage-controlled oscillator* (VCO). An example circuit, using a Clapp (series-tuned Colpitts) configuration is shown below:



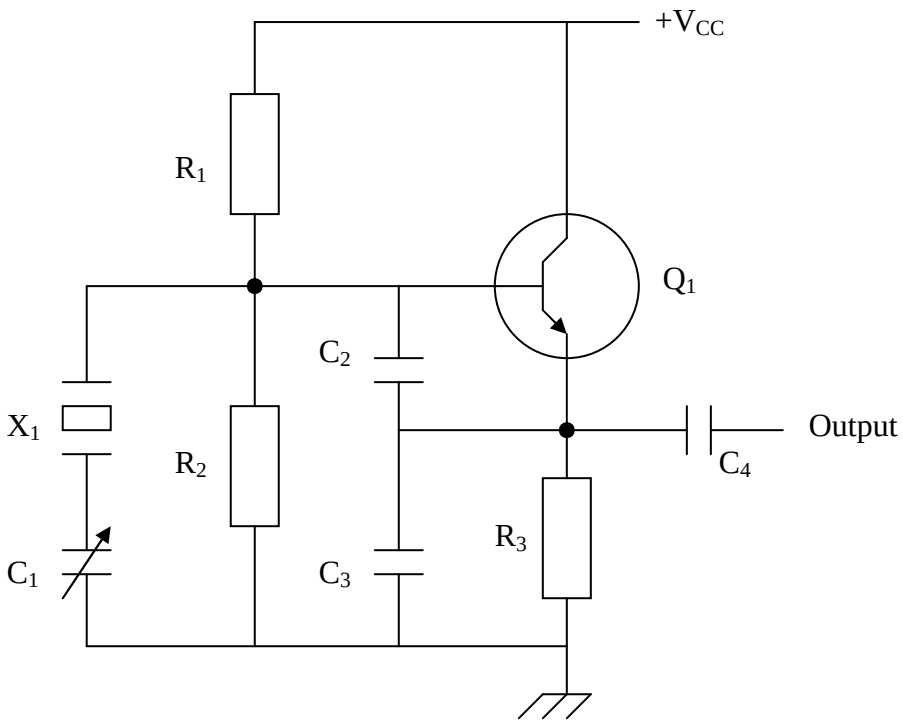
Voltage-Controlled Oscillator

The control voltage is applied through radio-frequency choke L_1 to reverse-bias the varicap diode D_1. This diode is in parallel with C_1, which provides some additional capacitance to supplement the typically low capacitance of a varicap diode. These capacitors are in series with L_2, hence the name "series-tuned Colpitts" oscillator (also called a Clapp oscillator). C_2 prevents the DC control voltage from interfering with the bias voltage generated by the voltage divider consisting of R_1 and R_2 (or *vice versa*). Q_1 is operated as a common collector (emitter follower) amplifier, and the output at the emitter of Q_1 is fed back into the tank circuit at the junction between C_3 and C_4, which form the tank circuit along with C_1, D_1 and L_2. The oscillator output is taken from the emitter of Q_1 via DC blocking capacitor C_5.

19.8 The Crystal Oscillator

Quartz crystals exhibit the piezoelectric effect. A voltage applied across the crystal causes the crystal to distort ("bend") slightly, and when the crystal returns to its undistorted shape a voltage is generated across it. As a result, the crystal appears similar to a series tuned

circuit and it can be used as the frequency-determining element in an oscillator. A typical circuit is shown below:



Crystal Oscillator

Here the resonant circuit consists of crystal X_1 with series capacitor C_1 and capacitors C_2 and C_3. Q_1 operates as a common-collector (emitter-follower) amplifier biased by R_1, R_2 and R_3. The output of the amplifier is fed back into the tank circuit at the junction between C_2 and C_3. This circuit also uses a series-tuned Colpitts or Clapp configuration.

Crystals have the advantage of providing very good frequency stability. The frequency of a crystal controlled oscillator will remain stable with little tendency to gradually change frequency (or "drift"). Drift is a problem with oscillators using traditional inductor-capacitor tuned circuits. The disadvantage of crystal oscillators is that they cannot be tuned over any great range. The variable capacitor C_1 in this circuit can vary the frequency slightly (which is known as "pulling" the crystal), but the tuning range is very limited. A crystal oscillator that allows the frequency to be varied is called a "variable crystal oscillator", abbreviated "VXO".

Summary

Oscillators are circuits that generate AC signals. Oscillators consist of an amplifier with positive feedback through a phase-shift network. The phase shift network usually also serves as a bandpass filter. An oscillator will oscillate at any frequency and amplitude where the Barkhausen criteria for oscillation are met:

- The loop gain is unity.
- The sum of the phase shifts around the feedback loop is zero or an integer multiple of 360˚.

The output of an oscillator should be buffered to prevent the frequency of the oscillator from changing as the load on the oscillator varies.

There are several different oscillator circuits, including the Colpitts, Hartley and Clapp oscillators, which differ in the precise arrangement of the tank circuit. An oscillator that allows the frequency to be varied is called a Variable Frequency Oscillator (VFO). If the frequency is varied by applying a control voltage, it is a Voltage Controlled Oscillator (VCO).

Quartz crystals exhibit the piezoelectric effect and act like series tuned circuits. They can be used to control the frequency of an oscillator. Crystal-controlled oscillators exhibit excellent frequency stability, with very little drift. However, they are essentially fixed-frequency oscillators. Although the frequency can be "pulled" slightly using a variable capacitor, the tuning range is not nearly as wide as for oscillators using ordinary tuned circuits. Crystal oscillators that allow the frequency to be varied are called "variable crystal oscillators", abbreviated "VXO".

Revision Questions

1 The names Clapp, Colpitts and Hartley refer to:

a. Transistors.
b. Power amplifiers.
c. Oscillators.
d. Diodes.

2 Which of the following is *not* a basic requisite for oscillation?

a. Feedback from output to input of the amplifier.
b. Correct phasing of input and output circuits.
c. Amplifying of signals from input to output.
d. Tuned circuit in both input and output stages.

3 The purpose of an amplifier in an oscillator is to:

a. Cancel phase shift.
b. Compensate for circuit losses.
c. Produce an increasing output.
d. Act as an oscillator buffer.

4 An oscillator varies its frequency as the loading on the following power amplifier is increased. In redesigning this circuit use should be made of:

a. A more powerful oscillator.
b. A well-regulated DC supply.
c. An intermediate buffer stage.
d. Decreased L/C ratio in the oscillator.

5 Colpitts, Clapp, Gouriet, Beat Frequency and Crystal are all types of:

a. Tuners.
b. Oscillators.
c. Antennas.
d. Amplifiers.

6 The characteristic of an oscillator which determines its operating frequency is:

a. Its resistance.
b. Its resonant frequency.
c. Its inductive reactance.
d. Its size.

7 The oscillator configuration where feedback is via a tapped inductor is:

a. The Armstrong oscillator
b. The Clapp oscillator
c. The Colpitts oscillator
d. The Hartley oscillator

8 A varicap diode might be used in an oscillator to:

a. Allow the frequency to be varied by a control voltage.
b. Regulate the supply voltage to the oscillator.
c. Limit the maximum amplitude of the output.
d. Rectify the output waveform to generate an automatic level control voltage.

9 At the frequency of oscillation, the loop gain of an oscillator is:

a. less than 1.
b. exactly 1.
c. greater than 1.
d. zero or an integer multiple of 360°.

10 Which amplifier configuration can be used in an oscillator?

a. Common base.
b. Common collector.
c. Common emitter.
d. Any of the above.

Chapter 20: Frequency Translation

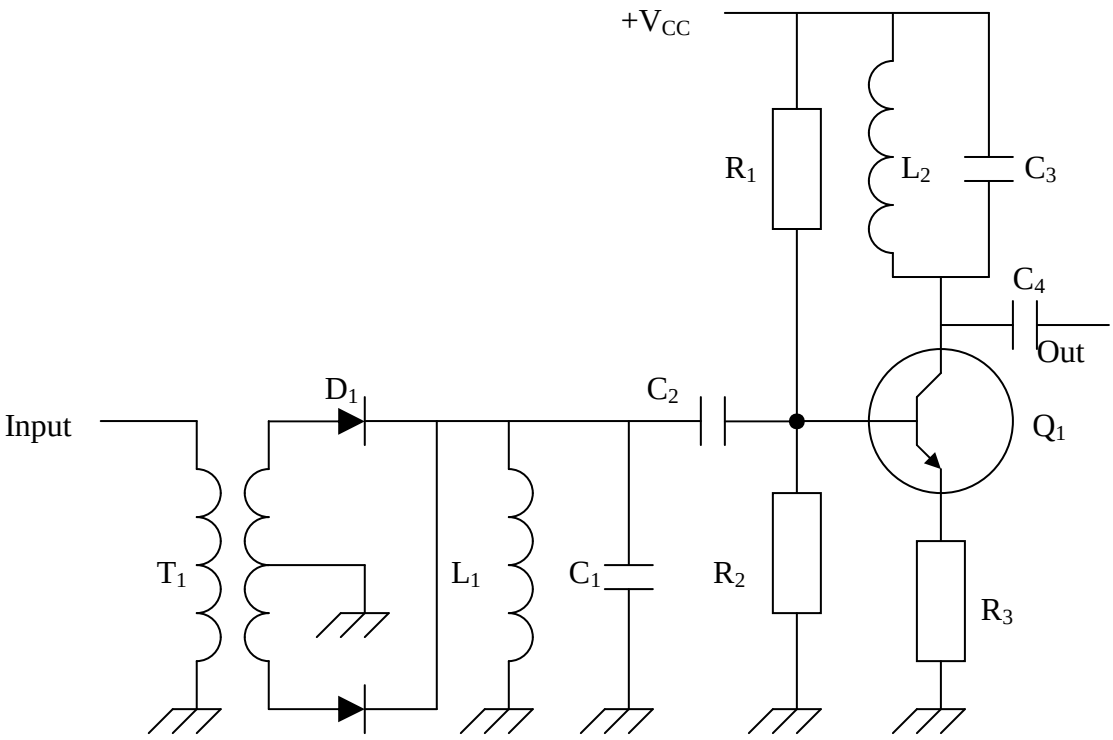
Oscillators are used to generate the signals of various frequencies that are needed by in transmitters and receivers. However, it is often useful to be able to create a signal of a desired frequency from signals of other frequencies. For example, it can be very beneficial to generate a signal at the precise output frequency the user has chosen from a very stable reference signal at a fixed frequency. The circuits we use to do this are frequency multipliers, dividers, frequency synthesisers and mixers.

20.1 The Frequency Multiplier

Any waveform that is not a perfect sine function contains harmonics as well as the fundamental frequency. The harmonic of a sine function is found at integral multiples of the fundamental frequency. For example, a 10 MHz signal that is not a perfect sine signal might have harmonics at 20, 30, 40, 50, 60 and 70 MHz, and so on. These harmonics are known as the second, third and further harmonics of the fundamental frequency (10 MHz in this case).

The presence of harmonics in a signal is can be used to create a frequency multiplier. The input sine signal is intentionally distorted; creating a signal that is rich in harmonics. The desired harmonic is then selected using a bandpass filter, yielding a signal that is some integer multiple of the input signal. The second and third harmonics are most commonly used. For example, a 7 MHz signal applied to the input of a 2x frequency multiplier would yield a 14 MHz signal; while 3x multiplier would yield a 21 MHz signal.

Different types of distortion result in different amounts of the various harmonics. When designing a frequency multiplier, the type of distortion introduced should maximise the desired harmonic. For example, a frequency doubler (a 2x multiplier) could use a full-wave rectifier to distort the input waveform, since the resulting rectified sine signal has a high second-harmonic content. A typical circuit is as follows:



Frequency Doubler

The input signal is full-wave rectified by T_1, D_1 and D_2. L_1 and C_1 form a parallel tuned circuit, which is resonant at the output frequency (twice the input frequency). It shorts the DC component of the full-wave rectified signal to ground and attenuates the undesired higher-order harmonics. Transistor Q_1, with resistors R_1, R_2 and R_3, forms a common-emitter amplifier. There is another parallel tuned circuit made up of L_2 and C_3 in the collector circuit of the amplifier, which further attenuates undesired high-order harmonics (3, 4, 5 times the input frequency etc.). C_2 and C_4 are DC blocking capacitors. The output is a sine signal at twice the frequency of the input signal.

A 3x multiplier (frequency tripler) might use a class C amplifier to introduce the necessary distortion, since the output of a class C amplifier has a high third-harmonic component. In VHF and UHF applications, varicap diodes are often used as the non-linear element to distort the input waveform and generate harmonics.

Because frequency multipliers introduce distortion, they cannot be used with signals that contain a range of frequencies, such as audio signals or amplitude modulated (AM) and single-sideband (SSB) RF signals. If they were, the many different frequency components of these signals would interact with each other, causing unwanted inter-modulation distortion (IMD) components, that are too close to the desired frequencies to be filtered out. However, they can safely be used with un-modulated signals, or with CW (Morse code), frequency modulated (FM) and phase modulated signals.

Frequency multipliers are only useful for multiplying by fairly small numbers, such as 2, 3 or 4. They cannot be used to multiply by large numbers – say 100 – because it would be too difficult to construct a filter to separate the 100^{th} harmonic from the 99^{th} or 101^{st} harmonics, and the nature of frequency multipliers means that they tend to generate at least some of most of the harmonics.

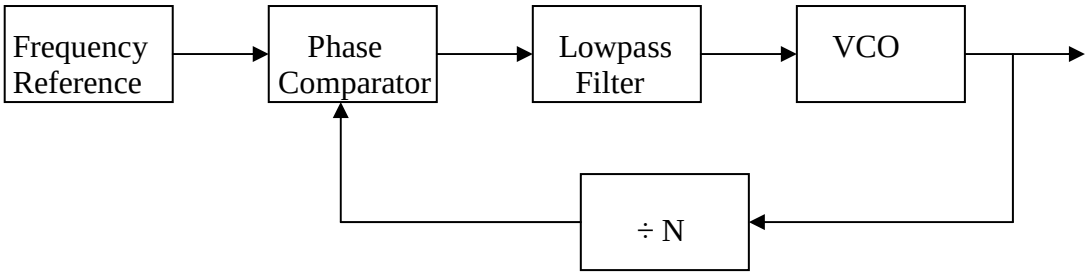
20.2 The Frequency Divider

Digital integrated circuits are available that can divide the frequency of an input waveform by any integer number – either a fixed number, or one that can be programmed by a microprocessor. The output of these "digital dividers" is typically a square wave, which contains high harmonic content (especially the odd harmonics at 3 times, 5 times, 7 times the input frequency and so on). This harmonic content can be removed using a suitable lowpass or bandpass filter, leaving a sine signal at the desired frequency.

20.3 The Phase Locked Loop Frequency Synthesiser

Although variable-frequency oscillators (VFO) can be used to generate a signal at a frequency selected by the user, they suffer the disadvantage that it is difficult to make them very stable. Their frequency tends to drift in response to changes in the ambient temperature, and to make more rapid excursions if bumped or otherwise maltreated. Crystal oscillators, on the other hand, are very stable in the face of temperature variations and mechanical knocks. However, their very limited tuning range makes them unsuitable for use as, say, the main oscillator for a transmitter that must cover an entire amateur band.

The most common solution in modern amateur equipment is to use a frequency synthesiser. A synthesiser is a circuit that can generate many programmable output frequencies based on a single reference frequency derived from a stable crystal oscillator. Although there are several different types of frequency synthesiser, this section will only cover one of these, the phase locked loop (PLL) frequency synthesiser. The block diagram of a simple PLL synthesiser is shown below:



Block Diagram of a PLL Frequency Synthesiser

The output of the frequency reference is fed into a phase comparer. This is a circuit that compares the phase of two signals and generates an output voltage that depends on the phase difference between the signals. This voltage is smoothed by a lowpass filter, and used to control the frequency of a voltage-controlled oscillator. The signal generated by the VCO is the input to a frequency divider that divides the input frequency by some (usually programmable) integer N. The output of the frequency divider is the second input to the phase comparer.

To understand how this circuit works, suppose that the frequency of the VCO is exactly N times the reference frequency. Then the phase comparer will generate a DC output voltage that is dependent on the phase difference between the two signals. This DC voltage will pass through the lowpass filter, and will affect the frequency of the VCO. Suppose the effect is to increase the frequency of the VCO slightly. As the frequency increases, the phase of the VCO output signal will begin to shift relative to the phase of the reference signal, which will change the output voltage of the phase comparer, which is the VCO control voltage.

The circuit is arranged so that if the frequency of the VCO increases slightly, the resulting output voltage from the phase comparer will reduce the frequency of the VCO again, to bring it back to its "correct" frequency, which is N times the reference frequency. Similarly,

if the frequency of the VCO decreases slightly, the resulting output voltage from the phase comparer will act to increase the frequency of the VCO, again returning it to a frequency of N times the reference frequency. In this condition, the VCO is said to be *phase locked* to the reference frequency, since any change in the phase relationship between the two signals (caused, for example, by a change in the VCO frequency) will act on the VCO in a way that will return it to the correct phase relationship with the reference frequency. This loop is an example of *negative feedback*

In case you were wondering, the reason for the lowpass filter is because most phase comparers actually generate a fairly complex output signal that has a DC (or low-frequency) component that reflects the phase difference between the inputs, as well as components at the different input frequencies to the phase comparer. We don't want the VCO to respond to all these little signal components. The lowpass filter rejects the high-frequency outputs, leaving only the low-frequency phase comparison voltage.

So now we have a circuit that can generate a frequency that is N times a stable reference frequency, and is phase locked to the reference frequency, so that it is almost as stable as the reference frequency itself. However, by changing the value of N, we can change the output frequency, making it any integer multiple of the reference frequency. If the reference frequency is small enough—say 10 Hz—then we can generate an output frequency that is any multiple of 10 Hz. For example, if the reference frequency is 10 Hz and the divider N is 1 402 000, the output frequency will be 10 x 1 402 000 = 14 020 000 Hz, or 14,020 MHz. If N is increased by 1 to 1 402 001, the output frequency would be 14 020 010 Hz. This arrangement allows us to synthesise almost any desired frequency from a single stable reference frequency. In modern radios, the divider N that controls the output frequency is usually set by a microprocessor in response to user input, such as adjusting the tuning dial.

The only remaining problem is to generate a stable 10 Hz reference frequency for our synthesiser. We can't use a crystal oscillator directly, since 10 Hz is much too low a frequency for a reasonably-sized quartz crystal. What we can do, though, is to run a crystal oscillator at a more suitable frequency – perhaps 100 kHz – and then use a digital divider to reduce the frequency to the desired reference frequency. In this case, dividing the 100 kHz oscillator output by a factor of 10 000 would give a 10 Hz reference frequency.

In practical PLL synthesisers it turns out that there is a trade-off between the speed at which the synthesiser can change its frequency (the "tuning rate", if you like) and the resolution of the synthesiser (its "step size"). The reason is that the resolution of the synthesiser is set by the reference frequency, so a high resolution requires a low reference frequency. But this low reference frequency requires a low cut-off frequency for the lowpass filter, which limits the speed at which the synthesiser can respond to changes in frequency. One solution is to combine the outputs of two synthesisers, one with a high reference frequency that can easily make large frequency changes but has limited resolution, and the other with a small step size that can "fill in" the missing frequencies, but which is never required to make large frequency changes (because the "coarse" synthesiser takes care of that). This arrangement is known as a *multiple-loop synthesiser*.

PLL synthesisers are very versatile and are the basis for most modern transceivers, allowing them to achieve very high stability combined with wide frequency coverage. However, they do have some disadvantages. In particular, early synthesisers such as those found in amateur equipment from the early 1980s suffered from significant phase noise, with the phase and frequency of the output signal varying very slightly as the loop adjusted it to keep it locked to the reference frequency. The result was phase noise that degraded the quality of both received and transmitted signals. Modern synthesiser designs are much better in this respect.

The phase noise from the synthesiser widens the spectrum of the VFO. Instead of a single line on the spectrum display, there is a single line with some fuzz at various frequencies above and below the primary frequency. Unfortunately, this fuzz influences both the receiver and the transmitter.

20.4 The Mixer

Another circuit that is commonly used for frequency translation in both transmitters and receivers is the *mixer*. It is based on the interesting mathematical result that if you multiply two sine signals together, you get a signal that consists of two components: one with a frequency that is the *sum of* the frequencies of the inputs, the other with a frequency that is the *difference between* the frequencies of the two inputs.

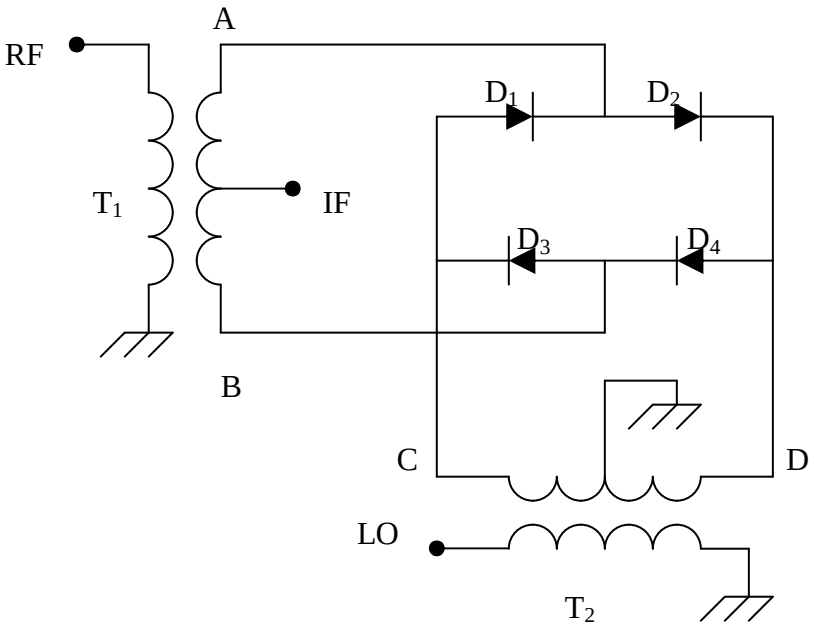
For the mathematically inclined, the relevant mathematical identity is:

$$2 \sin A \cos B = \sin (A+B) + \sin (A-B)$$

If you set $A = 2\pi f_1 t$ and $B = 2\pi f_2 t$ then the left-hand side, "2 sin A cos B" represents two sine signals of frequencies f_1 and f_2 multiplied together. The cos B introduces a phase difference, which is not relevant to this discussion. The right-hand side "sin(A+B) + sin(A-B)" represents the superposition (adding together) of sine signals with frequencies $f_1 + f_2$ and $f_1 - f_2$, the sum and difference frequencies.

For example, if you multiply a 9 MHz sine signal by a 6 MHz sine signal, you end up with two sine signals superimposed: one with a frequency of 15 MHz (the sum of the input frequencies), the other with a frequency of 3 MHz (the difference between the input frequencies). Electronic circuits that do this multiplication are called *mixers*.

Now it turns out that it is not very easy to accurately multiply two sine signals without introducing significant distortion into the output. One common solution is to use a switching mixer. Instead of actually multiplying the two signals together, it uses one of the input signals to switch the other input signal on and off, or to reverse its direction. Here is a typical circuit diagram for a switching mixer:



A Double-Balanced Diode Mixer

In this mixer, diodes are used as the switching elements. A strong input signal (generally derived from a *local oscillator*) is applied at the point marked LO, while a much weaker radio frequency signal is applied to the point marked RF. The output signal is taken from the point marked IF, for "intermediate frequency". The reason for these names will become apparent when we study the design of radio receivers.

The strong LO signal is used to "chop" the weaker RF signal, with the output appearing at the IF port. Here is how it works. Assume that the LO signals polarity is such that point C is positive with respect to point D. Diodes D_1 and D_2 will be forward-biased (turned on) while diodes D_3 and D_4 will be reverse-biased (turned off). If the diodes are properly balanced, with identical forward bias voltages, the point between D_1 and D_2 will be at the same potential as the centre-tap on the secondary winding of T_2, that is at chassis (earth) potential. Point A on the secondary winding of T_1 is at earth potential. If the polarity of the signal applied to the RF port is such that point A is positive with respect to point B, A will also be positive with respect to the output IF port, so the IF port will be negative with respect to point A, which as we have seen is at earth potential.

Now suppose the LO signal reverses polarity, while the RF signal remains as it was. Point D is now positive with respect to C, so diodes D_3 and D_4 will conduct, effectively earthing point B. Since the RF signal is making A positive with respect to B, it will also make the IF output positive with respect to B, which is earthed.

So in one half cycle of the LO (switching) input, the RF signal makes the IF output *negative* with respect to earth, while in the other half cycle, the RF signal makes the IF output *positive* with respect to earth. The result is that the LO signal is effectively switching the polarity of the RF signal as it appears at the IF output.

Hold on a moment. We started talking about *multiplying* two signals together, now we are talking about using one signal to switch the polarity of the other. What is the connection? It turns out that using the LO signal to switch the polarity of the RF signal is equivalent to multiplying the RF signal by a square function with the values +1 and –1. Multiplied by +1, the polarity is unchanged; multiplied by –1 it is reversed. One effect of this multiplication is that, because a square function contains not only the fundamental frequency, but also many harmonics, these harmonics are effectively mixed with the input signal as well. So instead of only getting the sum and difference frequencies, we also get the sum and difference frequencies of the RF signal and each harmonic of the LO signal. The unwanted mixing products can usually be filtered out by suitable filters following the mixer.

Diode mixers like this one require fairly high drive power at the LO port – typically +7 dBm (5 mW) or more. They usually exhibit a conversion loss of 6 to 7 dB, meaning that each of the output signals is 6 or 7 dB weaker than the RF input signal. However, they are widely used in amateur applications because they exhibit low distortion.

This mixer design is "double balanced" because neither the RF input signal nor the LO input signal will be reflected in the output. An unbalanced mixer would allow both the RF and the LO signals to get into the IF output, while a "single balanced" mixer would allow only one of these signals (typically the weaker and therefore less troublesome RF signal) to make it into the output.

There are many other mixer designs using transistors, specialised integrated circuits and other components.

One big advantage of mixers over other frequency translation circuits (frequency multipliers and the like) is that properly designed mixers do not introduce significant

distortion into the signals. They can be used with all types of signals, including amplitude modulated (AM), single sideband (SSB) and audio signals.

Summary

Frequency multipliers distort the input waveform to generate harmonics, and then select the desired harmonic using a bandpass filter. They can be used to multiply frequencies by small integers, typically 2 or 3. Frequency multipliers cannot be used with signals that contain many frequencies, such as AM or SSB signals, as they cause too much distortion. However, they can be used with CW and FM signals.

Digital integrated circuits are available that can divide a frequency by any integer number.

In a phase locked loop frequency synthesiser, the output frequency is locked to an integer multiple of a stable reference frequency. By changing the multiple, different frequencies can be generated from a single reference frequency. The output of a PLL synthesiser has similar stability to the reference frequency, although it will have additional phase noise.

The output of a mixer will contain signals with frequencies that are the sum of the frequencies of the input signals and the difference between the frequencies of the input signals. Depending on the mixer type, it may also contain signals at the same frequency as one or both of the input frequencies – if both input frequencies are suppressed then the mixer is "double balanced" while if only one input signal is suppressed it is "single balanced". Switching mixers will also typically contain mixing products caused by mixing various harmonics of the switching (LO) input with the low-level (RF) input. Unwanted mixing products must be removed by suitable filters at the output.

Revision Questions

1 A frequency multiplier stage is generally:

a. Biased into non-linearity.
b. Operated in class A.
c. Used with regeneration.
d. Used in processing SSB signals.

2 The circuit forming the basis of a frequency synthesiser is a:

a. Phase locked loop.
b. Automatic Gain Control.
c. Beat Frequency Oscillator.
d. Power Amplifier.

3 Frequency multiplication is often used in UHF transmitters. This is commonly achieved by applying RF power to diodes and tuned circuits. Such a device is a:

a. Varactor multiplier.
b. Heterodyne mixer.
c. Diode detector.
d. Power amplifier.

4 The reference frequency of a PLL frequency synthesiser is 10 Hz and the programmable divider is set to divide by 315 000. The synthesised frequency will be:

a. 315 kHz
b. 3,15 MHz
c. 31,5 MHz
d. 315 MHz

5 The cut-off frequency of the lowpass filter in a PLL frequency synthesiser will typically be:

a. Lower than the reference frequency
b. Higher than the reference frequency
c. Equal to the output frequency
d. Higher than the output frequency

6 A frequency multiplier could be used with the following signal without creating objectionable distortion

a. An amplitude modulated (AM) signal
b. A frequency modulated (FM) signal
c. A single sideband (SSB) signal
d. An audio-frequency voice signal

7 A local oscillator signal at 10 MHz is mixed with a 14 MHz signal, using a perfect double-balanced mixer. The output of the mixer will contain the following frequencies.

a. 10 MHz and 14 MHz only.
b. 4 MHz, 24 MHz and possibly other frequencies as well.
c. 4 MHz and 24 MHz only
d. 10 MHz, 14 MHz and 24 MHz only.

8 Which of the following circuits can be used to change the frequency of an amplitude modulated signal?

a. A frequency multiplier
b. A PLL frequency synthesiser
c. A mixer.
d. Any of the above

9 As well as the mixing products, the output of a single-balanced mixer will contain:

a. Nothing except harmonic mixing products.
b. One of the input signals.
c. Both of the input signals.
d. The average of the two input signals.

10 A switching mixer operates by

a. Reversing the polarity of one of the inputs depending on the polarity of the other.
b. Accurately multiplying two sine signals together.
c. Adding the two input signals together and then distorting the result to generate mixing products.
d. Relying on the square-law transfer characteristic of Field Effect Transistors.

Chapter 21: Modulation Methods

21.1 Modulation

Radio is based on the fact that electromagnetic waves of certain frequencies can travel great distances and still be strong enough to be detected by a radio receiver. However, in order for this effect to be useful, we need a way of sending information with, or imprinted upon, the radio waves. The sort of information that we wish to send—human speech, images or perhaps digital information—is not generally of the correct frequency to benefit directly from the ability of radio to span great distances. For example, the human voice has frequencies that range from approximately 50 Hz to 20 kHz. These frequencies are much too low to be effectively propagated as radio waves.

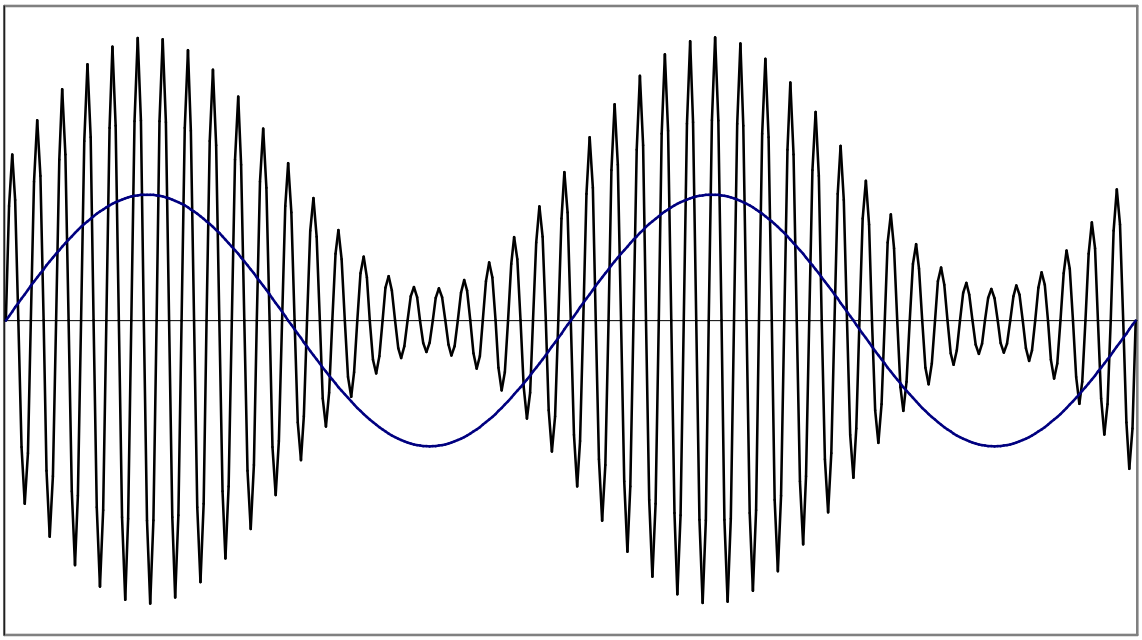
Modulation is the process of imprinting information on radio waves, so we can take advantage of the propagation of radio waves to transmit the information to a distant receiver. The radio wave is known as a *carrier*, while the information superimposed on the carrier is known as the *modulation* or *modulating signal*.

At the receiving end, the information must be retrieved from the carrier signal. This process is known as *demodulation* or *detection*. The original modulating signal (or at least a reasonable facsimile thereof) is the end product.

21.2 Amplitude Modulation (AM)

One of the earliest methods of modulation is *amplitude modulation*, or AM. Although not widely used in the amateur service any more, it still lives on in the AM transmissions of commercial radio stations in the medium frequency (or “medium wave”) broadcast band and in VHF aviation communications. In amplitude modulation, the amplitude (strength) of a radio frequency signal, called the *carrier* is varied according to the amplitude (strength) of the modulating signal.

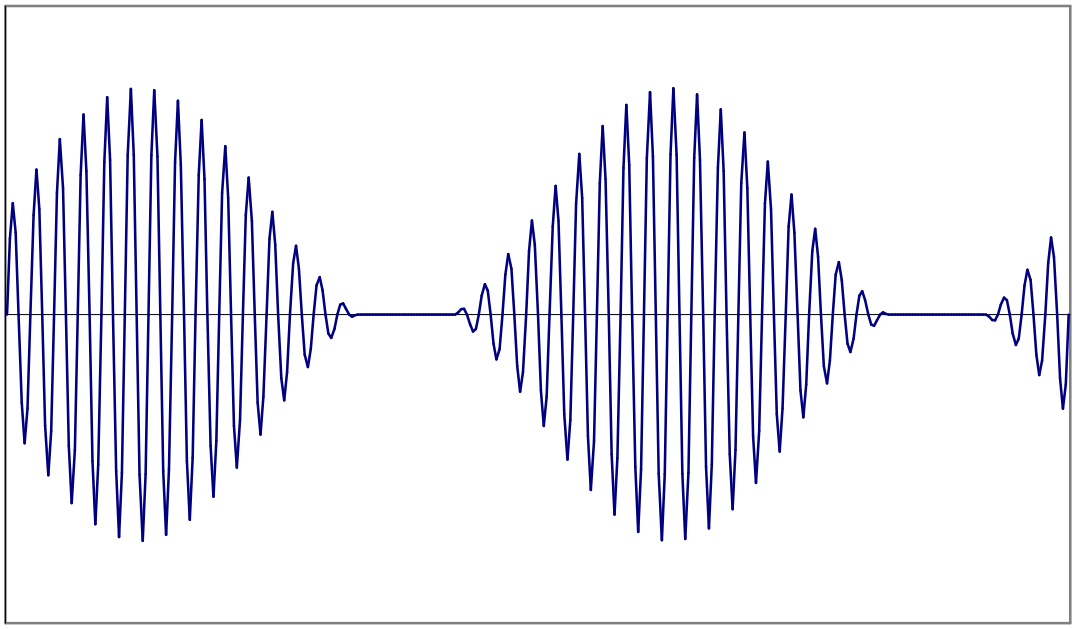
The plot below shows a low frequency sine signal, and the result when this signal is used to amplitude-modulate a higher frequency carrier.



A modulating signal and amplitude-modulated carrier

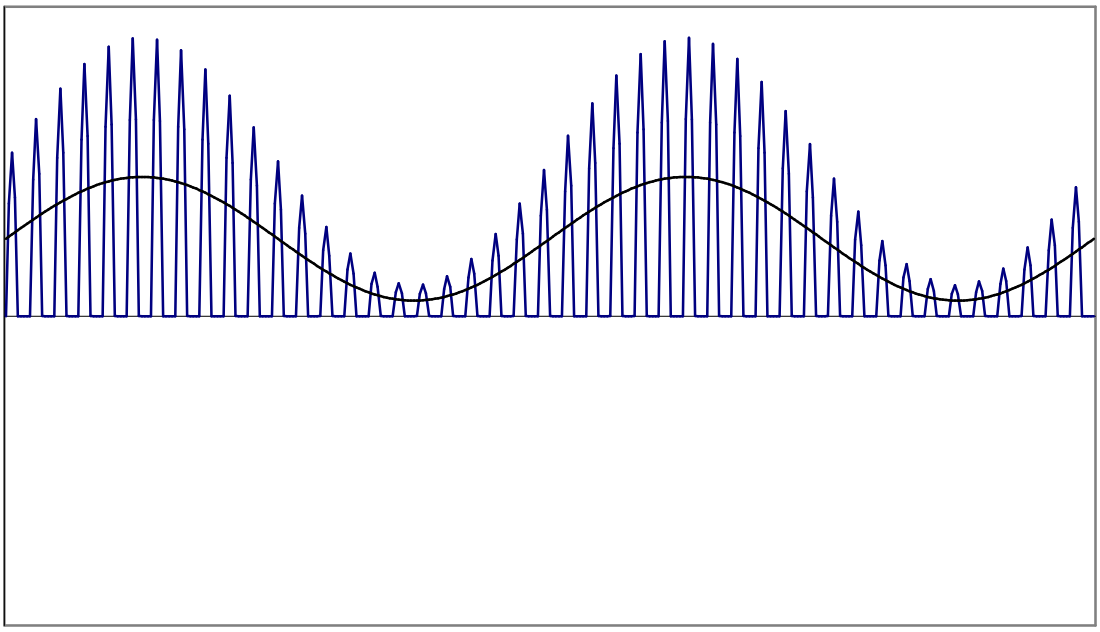
See how the amplitude of the high frequency carrier wave varies in step with the amplitude of the low frequency modulating signal. When the amplitude of the modulating signal is zero, the amplitude-modulated wave is at its “average” output level. When the amplitude of the modulating signal is positive, the amplitude-modulated signal is above this “average” amplitude, and when the modulating signal is below zero, the output is below this “average” level.

The *modulation depth* of an amplitude-modulated signal is the percentage by which the carrier signals varies above and below its average level in response to the modulating signal. In this example, the carrier is 80% modulated because the peak in the carrier amplitude is 80% above its average level, and the minimum carrier amplitude is 80% below its average level. The maximum possible modulation depth is 100% modulation. In a 100% modulated AM signal, the carrier amplitude decreases to zero when the modulating signal is at its most negative. Any attempt to modulate at more than 100% results in the carrier “bottoming out” at zero amplitude and distorting the modulation signal. This situation is known as *overmodulation* introduces a great deal of distortion. Because it causes interference to adjacent channels, overmodulation should be avoided. An example of an overmodulated signal is shown below:



An overmodulated AM signal

Amplitude modulation has the advantage that it is very simple to recover the modulating signal from the amplitude-modulated signal in the receiver. A simple half-wave rectifier followed by a lowpass filter will recover the modulating signal, which is typically an audio signal. The plot below shows a half-wave rectified AM signal, and the result of passing this signal through a lowpass filter.



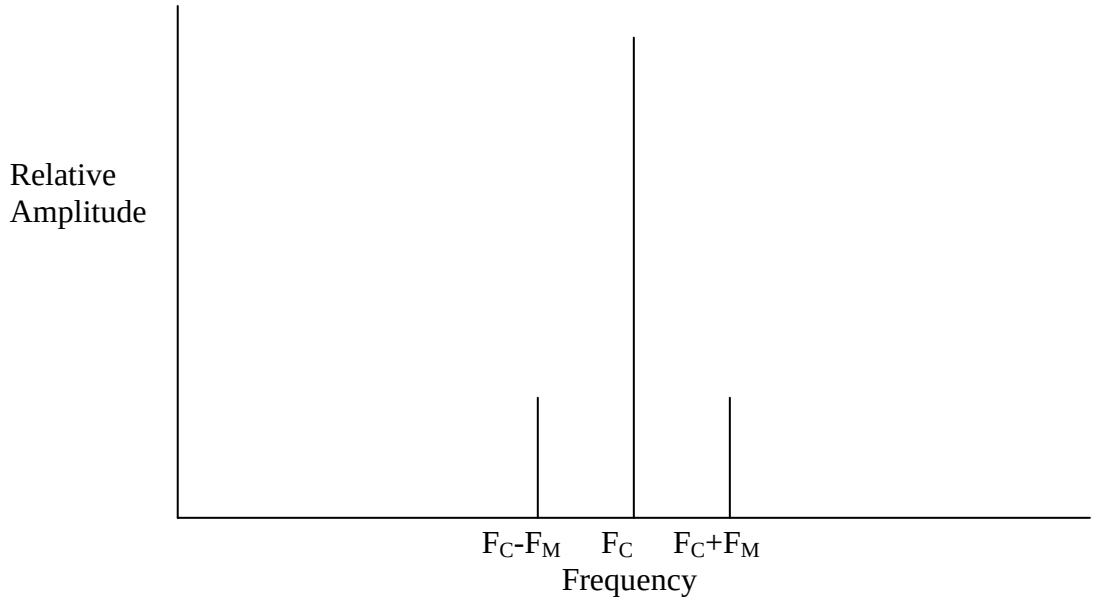
A half-wave rectified AM signal and the recovered modulation

The lowpass filter has removed what remains of the carrier, leaving the modulation "envelope" and a DC offset (which is indicated by the fact that the recovered signal is not symmetrical about the X axis). The DC offset can be simply removed by a blocking capacitor to obtain the original modulating signal. The process of recovering the modulation signal from a modulated signal is known as *demodulation* or *detection*.

Another way of looking at amplitude modulation is that it consists of multiplying the carrier by the modulating signal plus a DC offset. The value of the DC offset would be chosen to ensure that the sum of the modulating signal and the offset always remains positive, in order to prevent over-modulation. We know that multiplication implies a mixing process, so that amplitude modulation consists of *mixing* the carrier and modulation signals, resulting in an output that contains the *sum* and *difference* of the input frequencies, and possibly other components. In this case, the output also includes the carrier wave. The DC offset that we added to the modulating signal has a frequency of zero (because it's DC) which also mixes with the carrier, creating a sum frequency (the carrier frequency plus zero) and a difference frequency (the carrier frequency minus zero) that are both the same frequency as the carrier.

If the carrier frequency is F_C and the modulating frequency F_M, the amplitude-modulated signal will have frequency components of F_C, $F_C - F_M$ and $F_C + F_M$. These components can be plotted on a graph that shows frequency on the X-axis, and the relative amplitude of different components of the signal on the Y-axis. This graph is called the *frequency spectrum* of the signal.

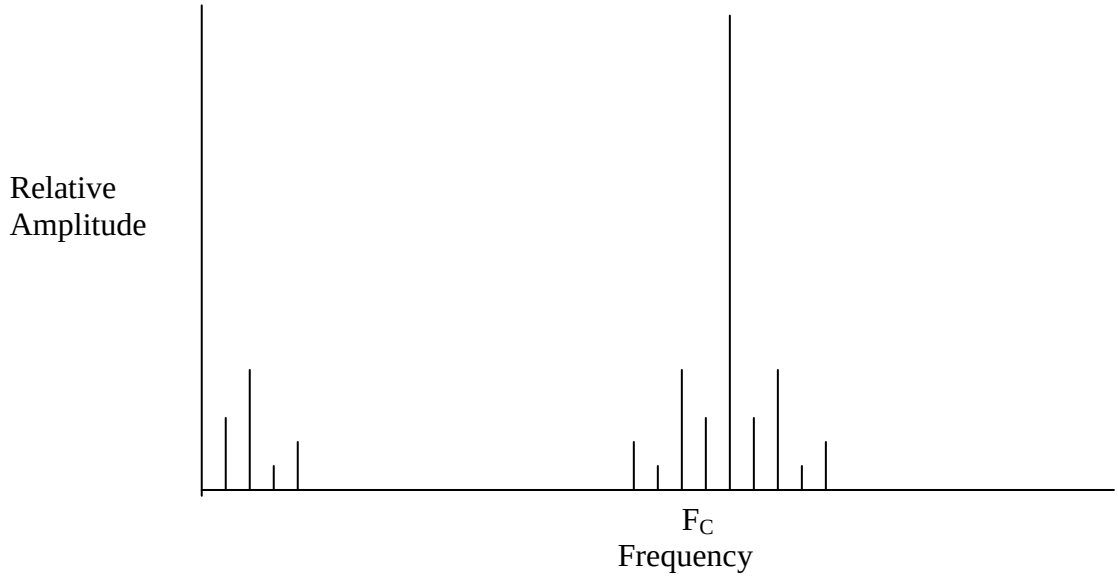
The vertical line above the carrier frequency F_C represents the carrier, while the lines above frequencies $F_C - F_M$ and $F_C + F_M$ represent the sum and difference frequencies respectively. Note that the carrier is much stronger than either of the other components. In an amplitude-modulated signal, two thirds of the power is contained in the carrier; the sum and difference frequencies together make up only one third of the total power of the modulated signal.



The frequency spectrum of a carrier amplitude-modulated by a sine signal

So far we have only considered a carrier that has been modulated by a single sine signal. However, speech consists of a whole range of frequencies, with many different frequency components present simultaneously in a speech signal.

Fortunately it is quite simple to figure out what happens if we amplitude-modulate a carrier with a speech signal that contains many different frequency components. Each of the different frequency components in the speech will create two output signals, one at the sum of the carrier frequency and this component of the modulating signal and one at the difference frequency. The following plot shows the frequency spectrum of some modulating signal (it is on the left of the graph, at a low frequency) and the corresponding amplitude-modulated signal. The modulating signal on this graph is called a *baseband* signal, as it is related to DC (a frequency of 0 Hz) just like the sidebands are related to the carrier.



Spectrum of a modulating signal and the corresponding amplitude-modulated signal

See how each component of the modulating signal corresponds to two components of the resulting amplitude-modulated signal, one above the carrier (the *sum*) and one below the carrier (the *difference*).

The total of all the "sum" components of the modulated signal – that is, all the components of the modulated signal that are higher in frequency than the carrier – is called the *upper sideband* of the AM signal. The total of all the "difference" components – that is, all the components of the modulated signal that are lower in frequency than the carrier – is called the *lower sideband* of the AM signal.

In order for speech to be reproduced intelligibly, frequencies from about 300 Hz to 3 kHz are required. This means that for a communications grade AM signal, such as is used in the amateur service, the upper sideband will extend from 300 Hz above the carrier to about 3 kHz above the carrier, while the lower sideband will extend from about 300 Hz below the carrier to about 3 kHz below the carrier. So the total *bandwidth* of the signal is 6 kHz, from 3 kHz below the carrier frequency to 3 kHz above the carrier frequency.

This analysis of the frequency spectrum of an AM signal shows the two greatest disadvantages of amplitude modulation.

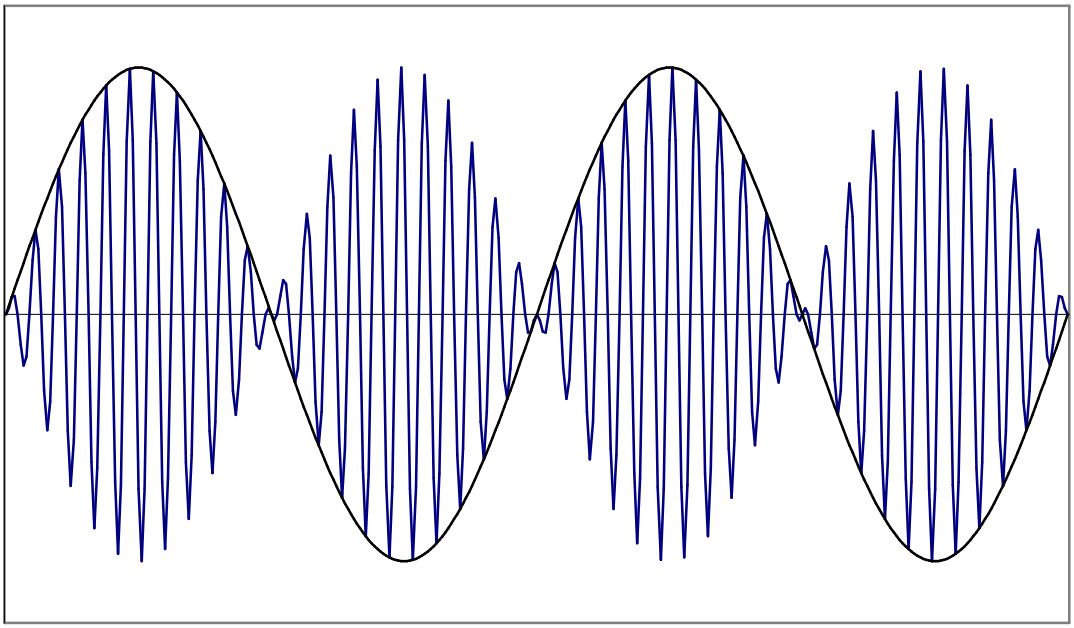
1. The component of the signal at the carrier frequency conveys no information (it is an unvarying carrier), and yet it consumes two thirds of the power of the signal. AM is therefore rather inefficient power-wise.

2. An AM signal transmits two copies of the modulating information, one in the upper sideband and one in the lower sideband, while only one of these sidebands would be sufficient to recover the original modulation. AM is therefore also rather inefficient in terms of the amount of spectrum (frequency spread) required. This effect is particularly important on the crowded amateur bands.

21.3 Double-Sideband Suppressed-Carrier Modulation

We could overcome the first of these problems – the power wasted by the carrier – if we could generate a signal without a carrier. This effect can be achieved by a *balanced modulator*, which outputs only the sum and difference components, but not the carrier itself. Mathematically this process is equivalent to simply multiplying the carrier signal by the modulating signal, without adding any DC offset. The plot below shows a low frequency sinusoidal modulating signal, and the resulting double-sideband suppressed-carrier modulated signal.

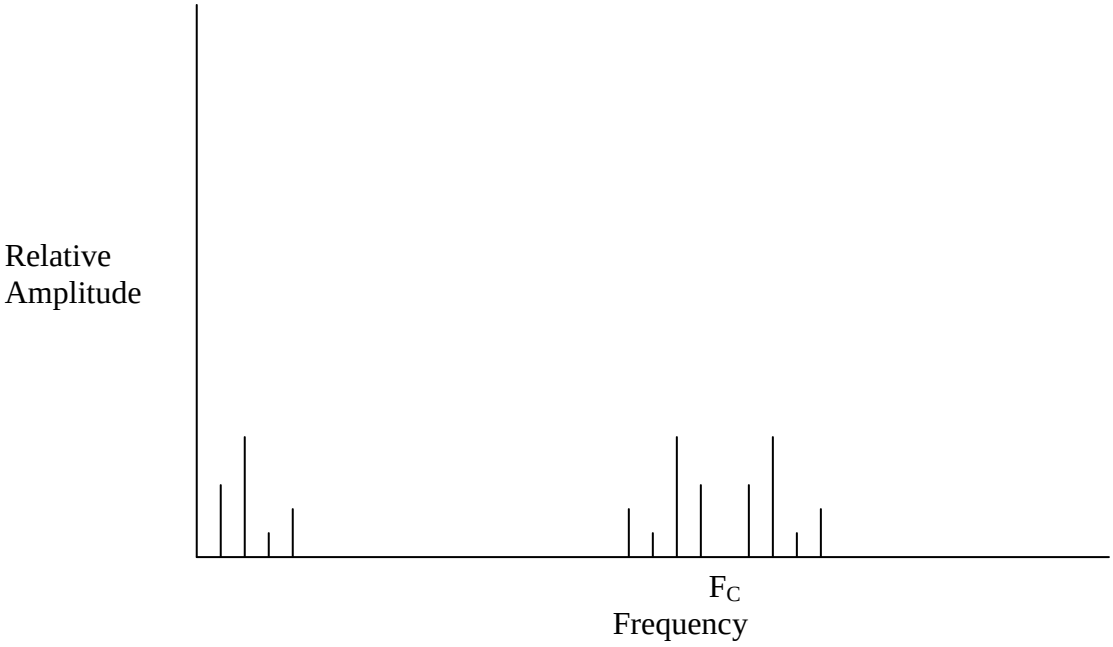
If you look carefully, you will notice that the phase of the modulated signal changes at the point where the modulating signal becomes negative.



A sine signal and double-sideband suppressed-carrier modulated signal

This time, because there is no DC offset on the modulating signal, the resulting double sideband modulated signal is zero when the modulating signal is zero. When the modulating signal goes from being positive to being negative or *vice versa*, the phase of the modulated signal is inverted, indicating that the modulating signal has crossed the axis. Note that you could not use a simple half-wave rectifier and lowpass filter to recover the modulation.

The frequency spectrum of a double-sideband, suppressed carrier signal is shown below, using the same multi-frequency modulating signal as in the last plot.

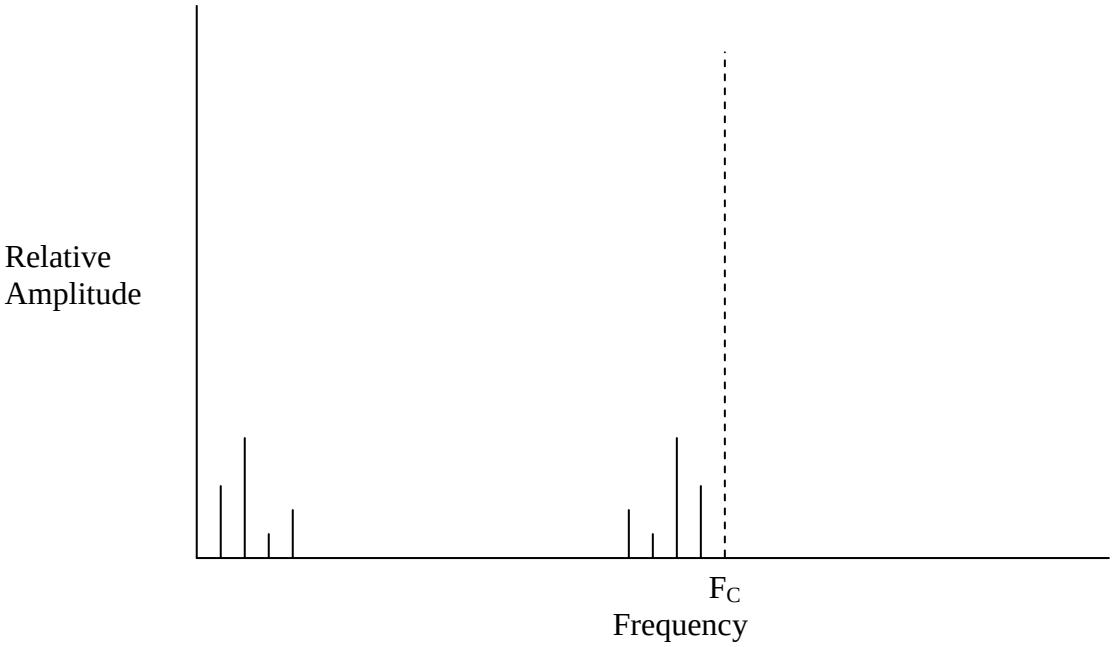


Modulating baseband signal and the corresponding double-sideband suppressed-carrier signal

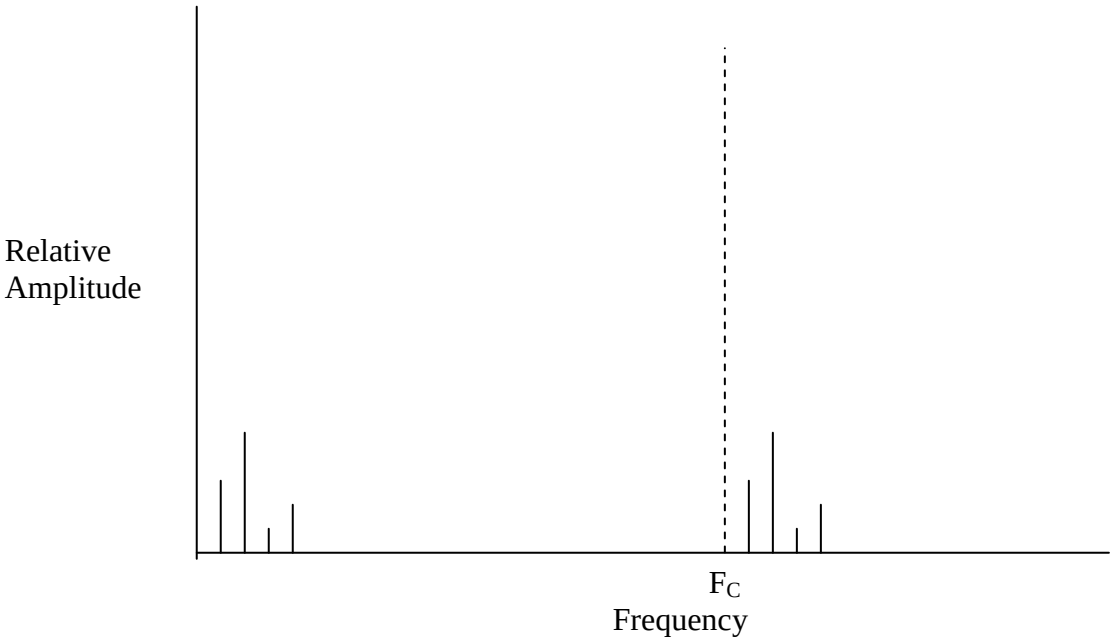
Not surprisingly, it looks exactly like the frequency spectrum of the amplitude-modulated signal, but without the carrier. Double-sideband suppressed-carrier signals are more power-efficient than amplitude-modulated signals, since they do not waste any power on the carrier. However, they still transmit the same information twice, wasting 50% of transmitted energy, and are as inefficient with spectrum usage as AM. For this reason, double-sideband suppressed-carrier signals are rarely used in practice.

21.4 Single-Sideband (SSB)

In order to avoid wasting bandwidth, we could simply take a double-sideband suppressed-carrier signal and remove one of the sidebands, leaving only a single sideband remaining. This type of modulation is formally known as "single sideband suppressed-carrier modulation", but is usually called just "single sideband" or "SSB". If we remove the lower sideband, the result would be an upper sideband (USB) signal. If we remove the upper sideband, the result would be a lower sideband (LSB) signal. The two plots below show the frequency spectra have lower sideband and upper sideband signals. The carrier frequency is shown as a dotted line so you can see where the frequency spectrum is in relation to where the carrier would have been if it had not been suppressed; but of course the carrier is not actually transmitted.



The frequency spectrum of a baseband modulating signal and the corresponding lower sideband signal



The frequency spectrum of a baseband modulating signal and the corresponding upper sideband signal

Note that in the lower sideband signal, the frequency spectrum of the modulating signal has been inverted (low frequencies in the modulating signal correspond to high frequencies in the lower sideband signal and *vice versa*), while in the upper sideband signal the spectrum in the modulated signal is the same way around as it was in the modulating signal. In fact, an upper sideband signal has an identical frequency spectrum to the original modulating signal, but translated to a higher frequency.

SSB is the most commonly used means of transmitting speech in the amateur service on the HF bands. Both upper and lower sideband are used. By convention, lower sideband is used on frequencies below 10 MHz, while upper sideband is used on frequencies above 10 MHz[4].

Because SSB signals do not have a carrier, the receiver frequency must be accurately adjusted to properly recover the original audio. Any maladjustment of the receiver frequency will result in the pitch of the audio being slightly too high or too low. Resulting signals can sound a little like Donald Duck (too high-pitched) or Darth Vader (too low-pitched). This distortion is not too important for speech, as it is easy to adjust the receive frequency sufficiently accurately to make speech intelligible, but it is the reason why AM or FM are usually preferred for music transmissions, where even a slight frequency shift in the received audio would be problematic.

21.5 Continuous Wave (CW)

Continuous Wave (CW) consists of turning the carrier on and off in order to convey information in Morse code. The name comes from the fact that the first transmitters used sparks, and were not capable of transmitting a continuous signal. Their transmitted signals would consist of an initial strong oscillation when the spark sparked that rapidly died down, known as “damped waves”. So when the first tube-based transmitters became available that were capable of transmitting continuously, they were called “continuous-wave” or “CW”

[4] The experimental 5 MHz band is an exception, with USB being used by convention.

transmitters, despite the fact that information was transmitted by turning the carrier on and off with the resulting combination of dits and dahs representing characters in Morse code.

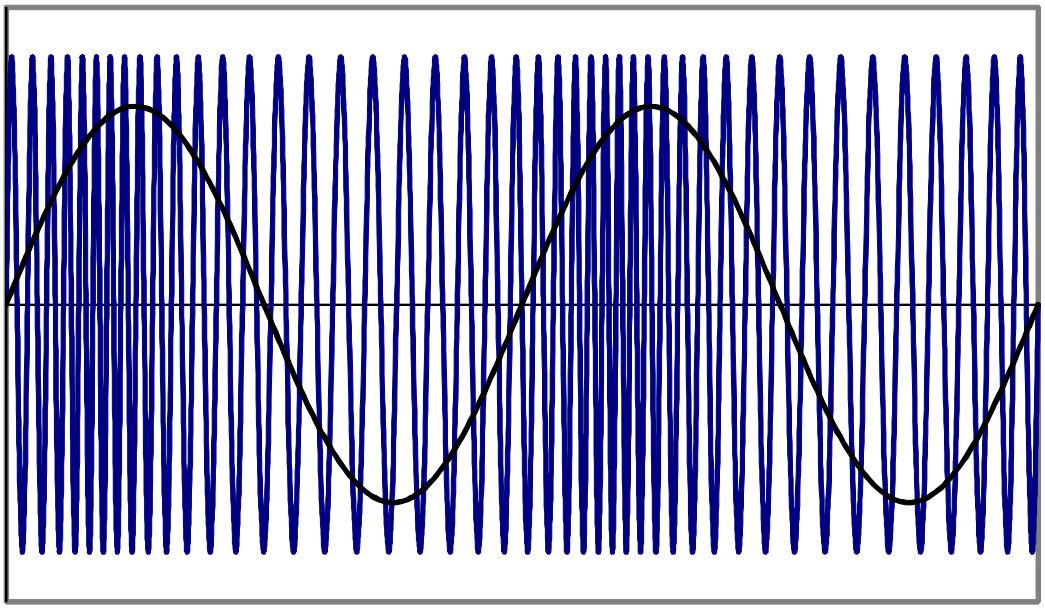
It might seem at first as though he frequency spectrum of a CW transmission should contain only the carrier, since the transmission consists of turning the carrier on and off. However, turning a carrier on and off is the same as amplitude-modulating it with a waveform that is at some fixed DC level when the carrier is to be turned on, or at zero when it is to be turned off, and so we should expect some sidebands in the keyed signal. Such sidebands do exist, but because Morse is normally sent at relatively slow speeds (lower than 50 Hz), the resulting bandwidth is quite narrow. When the keying waveform is an exact square function, the harmonic content of the keying waveform causes an increase in bandwidth, causing interference to adjacent users. This interference is called “key clicks” because clicks can be heard as clicks in a receiver when it is tuned close to, but not actually on the same frequency as, the CW transmission.

The shape of the envelope of the CW waveform – that is, the way it is turned on and off – has a big influence on the strength of the key clicks and how far they extend away from the carrier frequency. If the carrier reaches full amplitude as soon as it is turned “on”, and zero amplitude as soon as it is turned “off” then a lot of key clicks will be generated, causing noticeable interference to stations several kilohertz away. To avoid this interference, the carrier should be allowed to “ramp up” to full amplitude relatively slowly, and to decay back to zero over a while when it is turned off. The optimum ramp-up and decay period for a CW signal is around 5 ms. This waveform can be achieved using a capacitor that is charged and discharged through a resistor to determine the keying envelope. This circuit acts as a simple lowpass filter, attenuating the high-frequency harmonics of the keying waveform that would otherwise cause key clicks.

Although it may be branded as archaic, CW is still in widespread use. One of its advantages is that it is intelligible with much lower signal strengths than any voice signal. Practical listening tests have shown that CW requires about 13 dB less power for the same intelligibility as an SSB signal. A 100 W CW transmitter will “get out” as well as a 2 kW SSB transceiver. As of 2015, the major CW contests still attract tens of thousands of keen participants.

21.6 Frequency Modulation (FM)

Instead of varying the *amplitude* of the carrier depending on the amplitude of the modulating signal, frequency modulation (FM) varies the *frequency* of the carrier in response to changes in the amplitude of the modulating signal. For example, when the amplitude of the modulating signal is positive, the frequency might be increased slightly from the original carrier frequency, and when the modulating signal is negative, the frequency of the carrier might be reduced slightly. The plot shows a frequency-modulated signal:



A sine signal and the corresponding frequency-modulated signal

Note that the amplitude of the signal remains constant, while the frequency varies according to the amplitude of modulating signal. The amount of frequency change has been exaggerated to make it easier to see.

The amount that the frequency of the carrier increases or decreases in response to the modulation is called the *deviation* of the signal. The frequency of the carrier is both increased and decreased by the deviation, so for a signal with a deviation of 2,5 kHz, the frequency of the modulated signal will range from 2,5 kHz below the centre frequency to 2,5 kHz above the centre frequency, for a total change of 5 kHz.. The centre frequency is the carrier frequency with no modulation applied.

The *deviation ratio* or *deviation index* the maximum deviation divided by the highest modulating frequency.

$$m = \Delta F \div f_{mod}$$

For example, if the deviation is 2,5 kHz and the maximum modulating frequency is 3 kHz, the deviation ratio would be 2500 Hz/3000 Hz = 0,83.

The voice-grade FM transmissions typically used by amateurs are referred to as *narrow-band frequency modulation* (NBFM). In NBFM the deviation is kept to about 2,5 kHz and the resulting signal has a bandwidth of 5 to 6 kHz, comparable to that of a communications-grade AM signal. Commercial FM broadcast stations, by comparison, have a deviation of 75 kHz and a correspondingly much wider bandwidth. High deviation ratios trade bandwidth for quality. Although much spectrum space is used, the resulting clarity is ideal for broadcast applications, such as music stations.

FM signals have the advantage of better audio quality when the strength of the radio signal being received is fairly strong. When an FM signal is well above the atmospheric noise level, the amplitude variations due to noise have little effect on the receiver, which is only sensitive to variations in the signal frequency and not its amplitude. However, the quality of the recovered audio drops rapidly as the signal strength weakens and gets closer to the level of atmospheric noise. For this reason, amateurs mostly use FM for local communications on

VHF bands like the 2 m band (144 to 146 MHz) and UHF bands like the 70 cm band (430 to 440 MHz) where signals are usually strong and atmospheric noise is slight. For long-range communications in the high frequency (HF) bands between 3 and 30 MHz, where signals are often weak and atmospheric noise fairly strong, SSB is preferred.

21.7 Digital Modulation Techniques

So far we have concentrated on "human readable" signals, like the various phone (voice) modes and CW. However, an increasing role is being played by digital communications, where radio is used to transmit digital information between two computers. In this case, the information that is being transmitted consists of binary bits (ones and zeros). In fact, even speech signal are increasingly being transmitted in digital form, by first being converted into numbers, then transmitted, and then converted back into speech.

Frequency-Shift Keying (FSK)
A simple modulation method for digital information is frequency-shift keying, where the transmitter transmits one of two possible frequencies depending on whether it is sending a zero or a one. The two frequencies are called the "mark" and "space" frequencies, with the "mark" frequency corresponding to a logic "1" and the "space" frequency corresponding to logic "0".

FSK is used by modes such as RTTY (radio teletype), which allows interactive communication between two computers and Packet Radio, which provides electronic mail and file transfers over radio links. Amateurs started using RTTY with converted commercial teletype machines, but these days most RTTY is worked with a computer sound card. Modulation and demodulation are done in the computer, using software such as the free application MMTTY. Low baud rates of around 45 Bd (the unit "baud" means bits per second) are typically used.

Phase-Shift Keying (PSK)
Instead of shifting the *frequency* of the carrier, it is possible instead to shift the *phase* of the carrier depending on whether a one or a zero is being transmitted. The resulting modulation method is called phase-shift keying (PSK). PSK is preferred over FSK in most modern applications because it is more efficient in terms of bandwidth usage.

PSK comes in several different forms. In *binary phase-shift keying* (BPSK), the transmitted signal has one of two different phases, say 0˚ or 180˚, allowing one binary bit (a one or a zero) to be transmitted at a time. In *quadrature phase-shift keying* (QPSK), the transmitted signal can have one of four different phases (0˚, 90˚, 180˚ or 270˚), allowing two binary bits to be transmitted at a time.

The most popular amateur mode to use phase- shift keying is PSK-31, which is an interactive digital mode that allows two operators to "chat" to each other in real time over the radio. Everything that either operator types on his or her keyboard is immediately transmitted and displayed on the computer screen of the other operator (and anyone else who is listening). PSK-31 can use either BPSK or QPSK. When using QPSK the increased throughput is used to provide error detection and correction.

WSJT
Nobel physics prize laureate Joe Taylor K1JT has assembled a suite of different modulation modes in a single package called Weak Signal by Joe Taylor (WSJT). Many different modulation modes are included, all using variations of PSK and FSK plus multiple redundancy and very slow baud rates to achieve spectacular weak-signal performance. Using WSJT—which can be downloaded for free—and a normal sound card, anyone can enjoy weak-signal communications under conditions so poor that the operator may not even be able to hear a signal. The different modes are optimised for different applications.

JT65B, for example, is optimised for EME (earth-moon-earth) communications at 144 MHz, and has become the de facto standard for that mode. Using JT65B, EME is now within reach of almost anyone.

Detractors point out that their interest in ham radio is largely driven by what they hear. They wonder what the attraction is when your PC is contacting another PC, and the operator cannot even hear the weak signals…

Error Correction

Any modulation technique is prone to errors. Random noise can corrupt all or part of a transmission. It would be useful to know when such errors occur, and possibly even to correct them automatically.

Forward Error Correction (FEC) uses redundant information to allow the receiver to detect and possibly even correct errors, without having to communicate back to the transmitter. A simple FEC strategy would be to simply send each message twice. The two versions can then be compared at the receiver. If they are identical, they are probably correct. If they are not identical, there has been an error. Most FECs use an *error correcting code* (ECC), which contains enough redundant information inside each character to enable FEC. The first such code was Hamming's 7-4 code in the 1940, in which the five-bit codes were expanded to seven bits. Four of these bits had to be *1*. If they weren't, the character was not correct.

Retransmission protocols are used in packet-switched networks, in which the message to be transmitted is broken up into pieces known as *packets* before transmission. Any message consists of a number of packets, which can be passed individually through a network that may be unreliable. If the receiver receives a packet intact, it issues an acknowledgement (ACK). If the receiver detects that a packet has been damaged, it can request that the packet be re-sent (ARQ). If the transmitter receives an ARQ or does not receive an ACK after a certain period, it retransmits that packet. This strategy requires a two-way link and introduces considerable overhead, but produces error-free data. TCP/IP, which runs on the Internet, is an example of a retransmission protocol.

Summary

Modulation is the process of imprinting information on radio waves, so we can take advantage of the propagation of radio waves to transmit the information to a distant receiver. The radio wave is known as a *carrier*, while the information superimposed on the carrier is known as the *modulation* or *modulating signal*. At the receiving end, *demodulation* or *detection* results in the original modulating signal (or at least a reasonable facsimile thereof).

In amplitude modulation (AM), the amplitude of an RF carrier is varied according to the amplitude of the modulating signal. The resulting AM signal consists of the carrier, the upper sideband (at a higher frequency than the carrier) and the lower sideband (at a lower frequency than the carrier). The carrier takes two thirds of the power of an AM signal, with the remaining one-third of the power being shared equally between the upper and lower sidebands. Although AM signals are easy to demodulate using a half-wave rectifier and lowpass filter, they are inefficient both in terms of power (because the carrier conveys no information but takes 2/3 of the power) and bandwidth (since the modulating information is replicated in both sidebands).

Modulating signals are referred to *baseband* signals, and are generally found at low frequencies near DC (or 0 Hz).

A *double-sideband suppressed-carrier* signal is similar to an AM signal but without the carrier. It can be generated using a *balanced modulator*. The resulting signal is more power-efficient than an AM signal, but still uses twice the bandwidth of the modulating signal.

In a *single-sideband suppressed-carrier* (single sideband, or SSB) signal, the carrier and one sideband have been removed, leaving only a single sideband. SSB signals may be *upper sideband* (USB) or *lower sideband* (LSB). In LSB signals the spectrum of the modulating signal is inverted in the modulated signal; in USB, the spectrum is simply translated to a different frequency but is not inverted. SSB is one of the most efficient means of voice communications, especially when signal strengths are low.

Continuous Wave (CW) transmission consists of turning the carrier frequency on or off, and is used to send information in Morse code. CW is effectively a type of amplitude modulation, and the keying sidebands are known as "key clicks". Their extent and strength can be reduced by turning the carrier on and off more gradually, over a period of about 5 ms.

In *frequency modulation* (FM) the frequency of the carrier is varied according to the amplitude of the modulating signal while the amplitude remains constant. FM signals are capable of very good audio quality provided the received signal is fairly strong, but quality deteriorates rapidly as the received signal strength weakens. *Narrowband FM* transmissions by amateurs usually have a deviation of 2,5 kHz, resulting in a bandwidth of 5 to 6 kHz, which is similar to that of an AM transmission.

Frequency-shift keying (FSK) and *phase-shift keying* (PSK) are used to transmit digital information. In FSK, one of two frequencies is transmitted depending on whether a one or a zero is being sent; while in PSK the phase of the transmitted signal is varied to indicate that a one or a zero is being sent. FSK is used by modes like RTTY and Packet, while PSK is used by PSK-31. The WSJT suite is increasingly being used for weak-signal communications.

Forward Error Correction (FEC) and *retransmission protocols* introduce redundancy to detect and possibly correct errors in received data. Error-free data transmission can result, albeit at the expense of lower throughput.

Revision Questions

1 The process which alters the amplitude, phase or frequency of an radio wave for the purpose of conveying information is known as:

a. Alternation.
b. Microphonics.
c. Rectification.
d. Modulation.

2 The process of extracting information contained in an RF or IF signal is called:

a. Delination.
b. Degeneration.
c. Decoupling.
d. Demodulation.

3 The baseband signal is also known as the:

a. Carrier.
b. Modulating signal.
c. Upper sideband.
d. Lower sideband.

4 What does suppressing the carrier in an AM signal change the emission type to?

a. Single-sideband suppressed carrier.
b. Double-sideband suppressed carrier.
c. Frequency modulation.
d. Phase modulation.

5 What is one advantage of double-sideband suppressed-carrier transmission over standard full-carrier AM?

a. Only half the bandwidth is required for the same information content.
b. Greater modulation percentage is obtainable with lower distortion.
c. The transmitter is more energy-efficient.
d. Simpler equipment can be used to receive a double-sideband supressed-carrier signal.

6 A Class C frequency multiplier stage is unsuitable for raising the frequency of an SSB signal because of:

a. Impedance mismatch.
b. Severe distortion.
c. Lack of a carrier.
d. Untuned output circuits.

7 What signal component appears in the centre of an amplitude modulated transmitter's emitted bandwidth?

a. The lower sidebands.
b. The subcarrier.
c. The carrier.
d. The pilot tone.

8 In a frequency modulated signal, deviations from the carrier frequency depend on:

a. Amplitude of the audio signal.
b. Ratio of amplitude to frequency of the audio signal.
c. Frequency of the audio signal.
d. Frequency of the original carrier signal.

9 What sideband frequencies will be generated by an AM transmitter having a carrier frequency of 7250 kHz when it is modulated less than 100% by an 800 Hz pure sine signal?

a. 7250,8 kHz and 7251,6 kHz
b. 7250,0 kHz and 7250,8 kHz
c. 7249,2 kHz and 7250,8 kHz
d. 7248,4 kHz and 7249,2 kHz

10 The suppression of the carrier wave and one sideband in a transmission is known as:

a. Amplitude Modulation.
b. Frequency Modulation.
c. Single-sideband modulation.
d. Double-sideband modulation.

11 What determines the bandwidth occupied by each group of sideband frequencies generated by a correctly operating AM transmitter?

a. The audio frequencies used to modulate the transmitter.
b. The phase angle between the audio and radio frequencies being mixed.
c. The radio frequencies used in the transmitter's VFO.
d. The CW keying speed.

12 The term Narrow Band FM modulation usually refers to a signal of:

a. ± 2,5 kHz deviation.
b. 75 kHz deviation.
c. Low power levels.
d. Very stable frequency.

13 The bandwidth of a speech-quality AM transmission should not exceed:

a. 3 kHz
b. 6 kHz
c. 12 kHz
d. 24 kHz

14 When the modulation signal reduces the amplitude of a modulated wave to zero periodically, the modulation is:

a. 50 %
b. 100%
c. 200%
d. Overmodulated

15 The switching on and off of a transmitter to produce different lengths of carrier pulses for transmitting Morse code is called:

a. Current Injection.
b. Keying.
c. Demodulation.
d. Rectification.

16 CW, SSB, FM and AM are all types of:

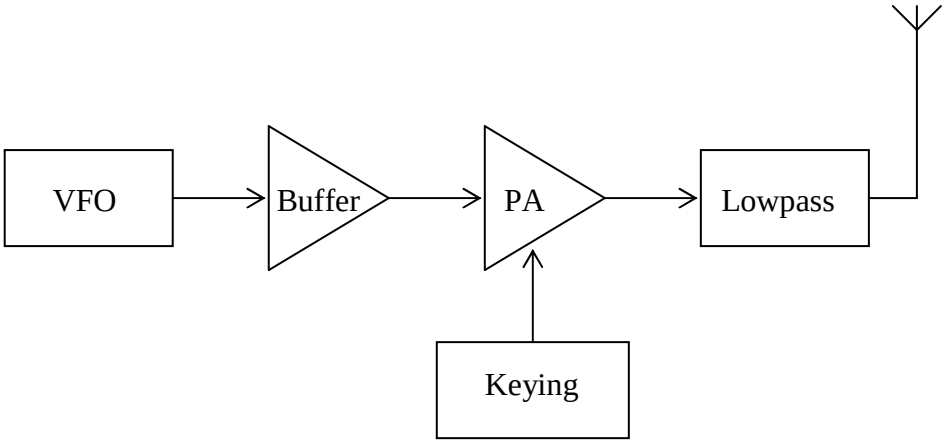
a. Time measurement.
b. Carrier modulation.
c. Radio Waves.
d. Amateur Licences.

Chapter 22: The Transmitter

The purpose of a transmitter is to generate a modulated radio-frequency signal that can be applied to an antenna. This module looks at the design of four typical transmitters: a single-band CW transmitter, a VFO-controlled AM transmitter, a simple SSB transmitter and a frequency-synthesised VHF FM transmitter.

22.1 A Single-Band CW Transmitter

One of the simplest transmitters is a VFO-controlled single-band CW transmitter. All you need is the variable frequency oscillator, a buffer amplifier (to prevent the variable loading of the power amplifier from affecting the oscillator frequency causing chirp), a keyed power amplifier and a lowpass filter to attenuate harmonics.



A Simple Single-Band CW Transmitter

In this design, the PA could run in Class C for maximum efficiency since linearity is not required when amplifying a CW signal. The non-linear amplifier would generate additional harmonics above the desired output frequency, but these could be easily eliminated by the output lowpass filter. The block labeled “keying” should include a keying waveform shaper, to prevent the key-clicks that would be caused by turning the carrier on or off too rapidly.

A design like this would be most suitable for the 80 m (3,5 MHz) or 40 m (7 MHz) bands, to keep the VFO frequency fairly low in order to allow reasonable frequency stability. VFOs are usually best kept below 10 MHz for good stability.

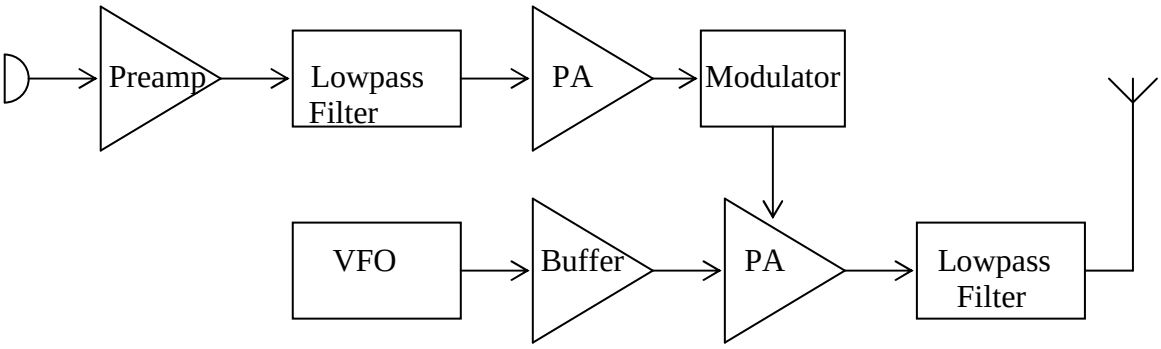
22.2 An Amplitude-Modulated (AM) Transmitter

There are two different ways to build AM transmitters. One way is to generate a low-level amplitude-modulated signal, and then amplify this signal to obtain the desired output power. This approach has the disadvantage that linear amplification is required because the AM signal contains many frequency components and non-linear amplification would cause inter-modulation distortion. However, it is the most common method in modern *multimode* transceivers that can generate AM, SSB and CW signals (and often also FM). This is because low-level modulation is the simplest way to generate an SSB signal, and the same circuitry can also be used to generate an AM signal.

However, for specialised AM transmitters there is an alternative, which is to generate the carrier signal and amplify it up to the desired output power, and then use a high-level modulator to modulate it at the full output power. This technique allows more efficient Class C amplifiers to be used to amplify the carrier signal, since before it is modulated it

contains only a single frequency component (the carrier frequency) and so does not suffer from inter-modulation distortion.

The following circuit shows a VFO-controlled AM transceiver using high-level modulation.

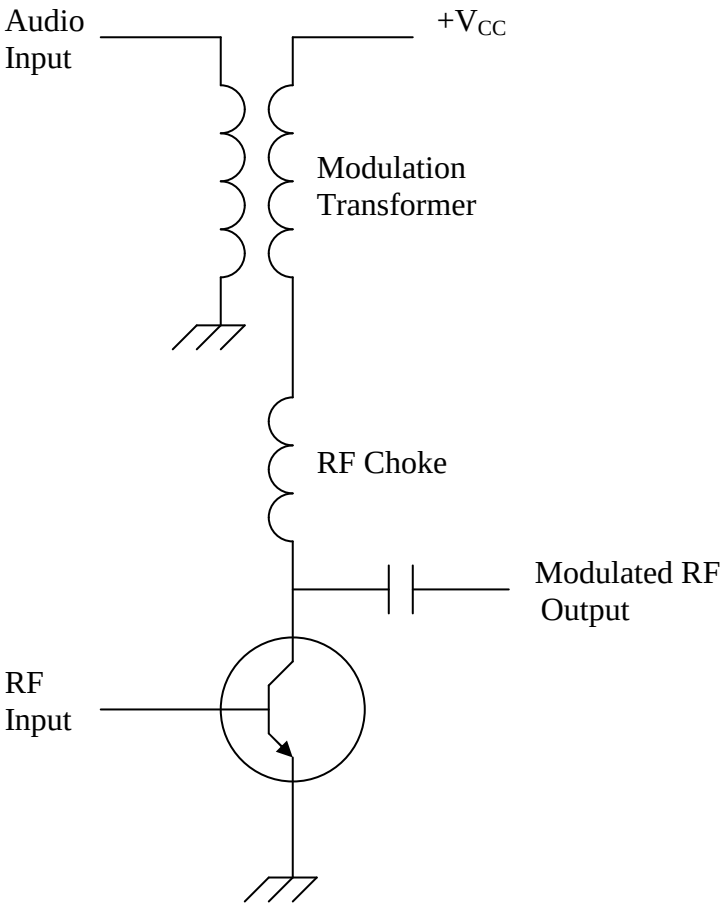


A VFO-controlled AM transmitter using high-level modulation

The audio input from the microphone is pre-amplified and then filtered to remove audio components above the voice range of 300 Hz to 3 kHz, required for communications-grade speech. The audio signal is further amplified by a power amplifier and fed to a high-level modulator that controls the Class C RF power amplifier. The input to this amplifier comes from a VFO operating on the intended output frequency.

Two-thirds of the energy in an amplitude-modulated signal is contained in the carrier and the remaining one-third in the modulation sidebands. In this circuit, the energy for the modulation sidebands is provided by the audio power amplifier. So if the carrier power were 100 W, the audio power amplifier would have to supply 50 W to fully modulate the signal.

A high-level modulator typically consists of a *modulation transformer* that modulates the supply voltage to the final output stage depending on the audio modulation. An illustrative circuit diagram is shown below.

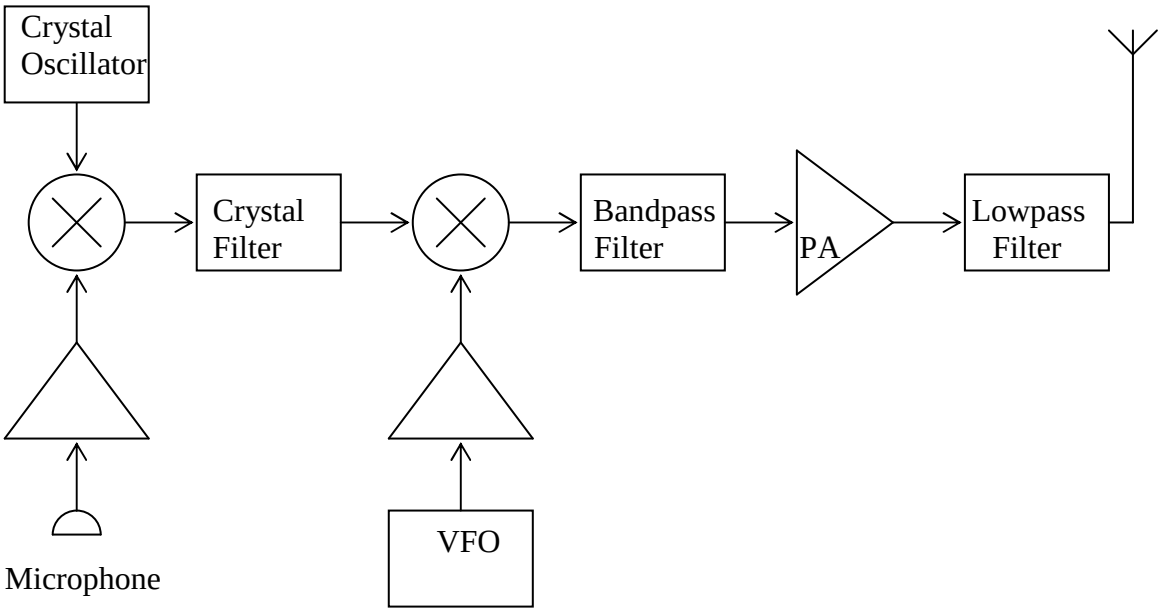


High-level modulation using a modulation transformer

For an example of an AM transmitter using low-level modulation, see the simple SSB transmitter described below. If the balanced modulator is replaced with an unbalanced modulator, and a crystal filter is used that is wide enough to permit both the upper and the lower sidebands to pass, the result is a low-level modulated AM transmitter.

22.3 A Simple SSB Transmitter

The following block diagram shows a simple single-band VFO controlled SSB transmitter for the phone segment of the 20 m band, from 14,100 to 14,350 MHz.



A Simple SSB Transmitter

In this simple SSB transmitter, the carrier is generated by a crystal oscillator at a fixed frequency, perhaps 9,000 MHz. This carrier is modulated by the amplified audio input in a balanced modulator (represented here by a circle with a cross inside it, the symbol for a mixer). Because the modulator is balanced, the output signal contains the upper and lower sidebands, but no carrier (so it is a double-sideband suppressed-carrier signal). A very narrow bandpass crystal filter is used to select the upper sideband only, i.e. frequencies from 9,0003 to 9,0030 MHz, eliminating the lower sideband. This technique is called the "*filter method*" of SSB generation.

Note that most filters that are sufficiently selective to pass one sideband while rejecting the other are fixed-tuned[5], so the resonant frequency cannot be altered. The SSB signal must be generated at a fixed frequency and then mixed up or down to the desired output frequency.

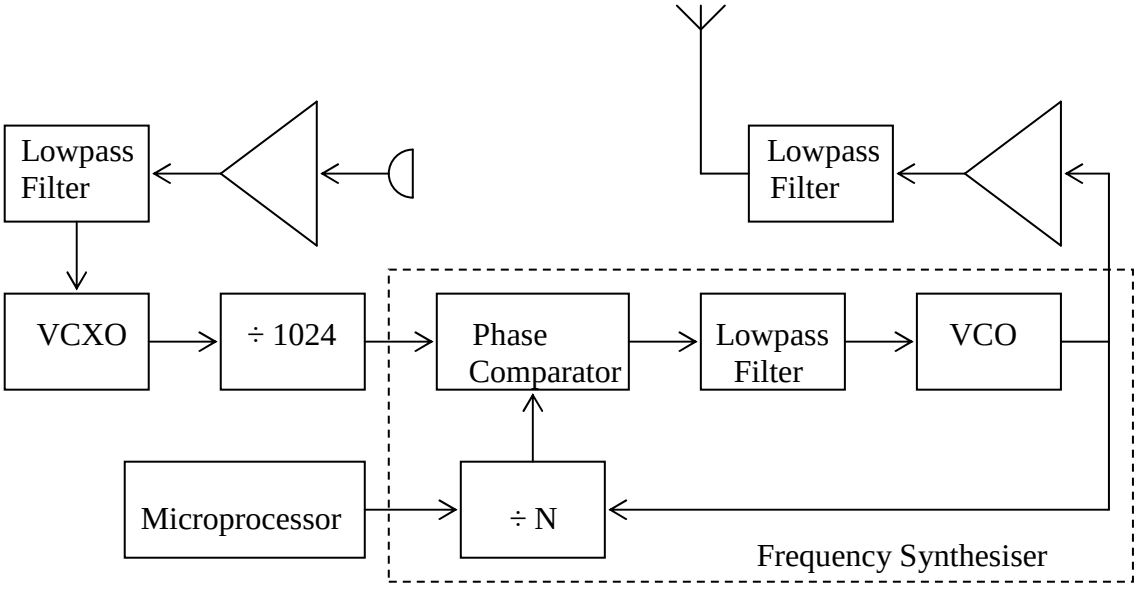
In this case the 9 MHz upper sideband signal is mixed with the output of a variable-frequency oscillator that ranges from 5,100 to 5,350 MHz, resulting in two signals. The sum will be a USB signal in the range 14,100 to 14,350 MHz, while the difference will be an USB signal ranging from 3,900 to 3,650 MHz (inverted). The bandpass filter following the mixer is an ordinary inductor-capacitor filter, which is designed to pass the frequency range 14,100 to 14,350 MHz (the phone segment of the 20 m amateur band) while rejecting frequencies in the range 5,100 to 5,350 MHz, the unwanted mixing product.

The filter is followed by a linear RF power amplifier (probably running in class AB) and a final lowpass filter that will pass the desired output frequencies in the range 14,100 to 14,350 MHz while rejecting harmonics at 28,200 MHz and above.

[5] With increasing use of DSP techniques, this statement will soon cease to be true.

22.4 A Frequency-Synthesised VHF FM Transmitter

Frequency synthesis is a natural approach for building a VHF FM transmitter, since it is not possible to make a VFO run with sufficient stability at VHF frequencies. Also, since most FM operation takes place at distinct frequency "channels" spaced 12,5 or 25 kHz apart, a simple single-loop synthesiser will suffice.



A Synthesised VHF FM Transceiver

The signal from the microphone is amplified and then filtered to restrict it to the communications voice range of frequencies below 3 kHz. The signal then frequency-modulates a voltage-controlled crystal oscillator (VCXO) running at 12,8 MHz, which would probably use a varicap diode to "pull" the crystal frequency slightly. The frequency-modulated output of the crystal oscillator is then divided by 1024 to generate a frequency-modulated 12,5 kHz reference signal for the PLL frequency synthesiser, which is made up of the functional blocks shown in the dashed rectangle.

Since the voltage controlled oscillator (VCO) in the frequency synthesiser is phase-locked to the reference frequency, it will follow the slight changes to the reference frequency caused by the frequency modulation, so the output of the frequency synthesiser will also be frequency-modulated. To cover the entire 2 m band the "÷N" divider in the frequency synthesiser would range from 11 521 (for an output frequency of 144,012 5 MHz) to 11 679 (for an output frequency of 145,987 5 MHz). The ÷N divider would be controlled by a microprocessor, which would select the correct division ratio according to the frequency set by the user.

The output of the frequency synthesiser would be amplified by the power amplifier and filtered by a lowpass filter to remove harmonics.

Chapter 23: Receiver Fundamentals

A radio receiver is the heart of any amateur radio installation, whether it is a stand-alone receiver or combined with a transmitter as a *transceiver*. It is relatively easy to build a good transmitter. All you really need is good frequency stability, adequate power and a clean output signal (no harmonics, key clicks or inter-modulation distortion). It is much harder to build a good receiver, and consequently there is more variation in receiver capability amongst both commercial and homebuilt designs.

23.1 Noise in Receivers

The main limitation to a receiver's ability to demodulate a signal is the *signal to noise ratio* (SNR). This ratio is normally expressed as the ratio (in dB) of the total signal entering the antenna terminals (including noise) to the noise itself. So:

$$SNR = (S + N) \div N$$

S is the desired signal; *N* is the noise, including anything except the desired signal. Other signals are also noise, for our purposes.

The following sources of noise contribute to the term *N* in the above equation:

- **Receiver thermal noise:** Semiconductors produce noise due to the semi-random movement of electrons in the semiconductor material. This noise is temperature dependent. The intensity of the noise is expressed as $P_N = kTB$, with P_N being the noise power, k being Boltzmann's constant, T being the noise temperature in K and B being the bandwidth in Hz. The higher the temperature, the more noise. The more bandwidth, the more noise. In some specialised applications, receiver components are actually cooled down to reduce receiver noise. However, specialised semiconductors can produce lower noise temperatures in the receiver. Good design and construction practices can reduce the noise in the receiver, leading to a lower effective temperature.
- **Other receiver noise:** Early synthesisers created lots of phase noise. This noise was due to phase jitter in the PLL's VCO. This noise also contributes to receiver noise, and is bandwidth dependent, just like thermal noise.
- **Band noise:**
  - **Atmospheric noise:** Distant thunderstorms contribute to a general noise level on the band, which masks weak signals. Atmospheric noise is mostly a problem on the low bands, except when there is heavy weather close by.
  - **Electrical noise:** Powerlines, switching gear, automotive ignition systems and electric motors may produce noise that enters the antenna.
  - **Ground noise:** The ground around the antenna radiates noise, which is temperature dependent, much like semiconductor noise.
  - **Galactic noise:** Certain galaxies and regions in the sky radiate lots of noise. The sun is the most dominant of these sources. In general, galactic noise is only a problem at VHF and above, as the noise otherwise does not penetrate the ionosphere below the critical frequency.
  - **Other signals:** Any signal except the one that we specifically want to receive contributes to noise.

At HF and below, the band noise is normally well above the receiver noise. However, at VHF and above, the receiver noise starts becoming the limiting factor. Designers have to go to great lengths to reduce the receiver noise to be able to hear weak signals. At these frequencies, low noise preamplifiers, low-loss cables and even cryogenic cooling come into play.

The noise figure of a receiver is specified in dB. The same information can be stated as a *noise temperature*. The advantage of this approach is that the simple formula $P_N = kTB$ can be used to calculate the actual noise power in the receiver, which can then be simply compared to the incoming signal to determine SNR. The temperature is stated in kelvin (K), which has the same magnitude as °C but starts at a different point. At 0 K, there is no kinetic energy due to temperature. It is also known as *absolute zero*. At absolute zero, there is no thermal noise. 0°C = 273 K, and 100°C = 373 K. Normal room temperature is therefore around 300 K.

23.2 Receiver Characteristics

Selectivity

When propagation conditions are good (i.e. strong radio signals are propagating long distances) the amateur bands can be a very crowded place. If you listen during any CW contest, for instance, you will hear signals spaced 100 to 200 Hz apart over the entire CW section of a band. So the first attribute a good receiver must have is *selectivity*, the ability to distinguish between close-spaced signals and receive only the one that the listener is interested in. The selectivity is mostly determined by bandwidth, specified in Hz.

Sensitivity

Many of the signals on amateur bands are very weak, having come from low-powered transmitters a long distance away, so the second attribute an amateur receiver needs is *sensitivity*, the ability to "hear" very weak signals. Sensitivity can be expressed in terms of the voltage (in µV) or power (in dBm) to produce a specific SNR.

Dynamic range

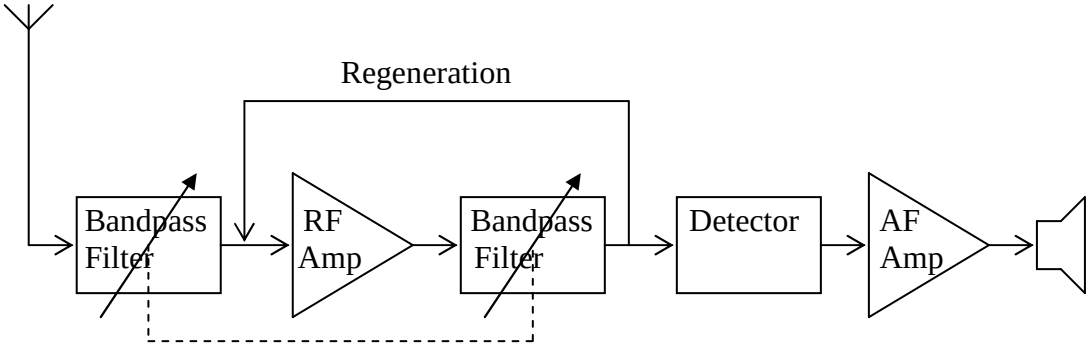
Since these weak signals may be adjacent to strong signals, perhaps from other amateurs in your town, amateur receivers need another attribute: *dynamic range*. Dynamic range is the ability to receive a weak signal despite the presence of other much stronger signals on nearby frequencies, and is expressed as a ratio in dB, along with the separation between the signals. The further apart the signals are, the more interfering signal the receiver can tolerate.

To get an idea of the challenges faced by receiver designers, a typical weak signal on an amateur band might deliver a power of –120 dBm from the antenna—that's one thousand-millionth of a microwatt. A strong signal might deliver –30 dBm, or one microwatt. So a strong signal could be 90 dB (a thousand million times) as strong as a weak signal—and yet the receiver might need to select and amplify the weak signal to a usable level, without being affected by the strong signal a few kilohertz away!

This module introduces two simple receiver designs—the *tuned radio frequency* receiver and the *direct-conversion* receiver—and considers how well they meet these requirements. It also introduces many of the concepts that you will need for the next module, which covers the *superheterodyne receiver*.

23.3 The Tuned Radio Frequency (TRF) Receiver

One of the simplest receiver designs, which has been with us almost since the dawn of radio, is the *tuned radio frequency* receiver. The principle is simple: you use a bandpass filter to select the signal you want, amplify the weak radio signals, demodulate the signal (to recover the audio modulating frequency) and then amplify the recovered audio sufficiently to make it audible in headphones or a loudspeaker. The block diagram below shows the layout of a TRF receiver. The block labeled "detector" is a half-wave rectifier to demodulate AM signals,



A Tuned Radio Frequency Receiver with Regeneration

The arrows through the bandpass filter indicate that they are tunable, so they can be used to select the desired signal. The dotted line joining the arrows on the two bandpass filters mean that they tune *together*, so a single control will change the tuning of both filters.

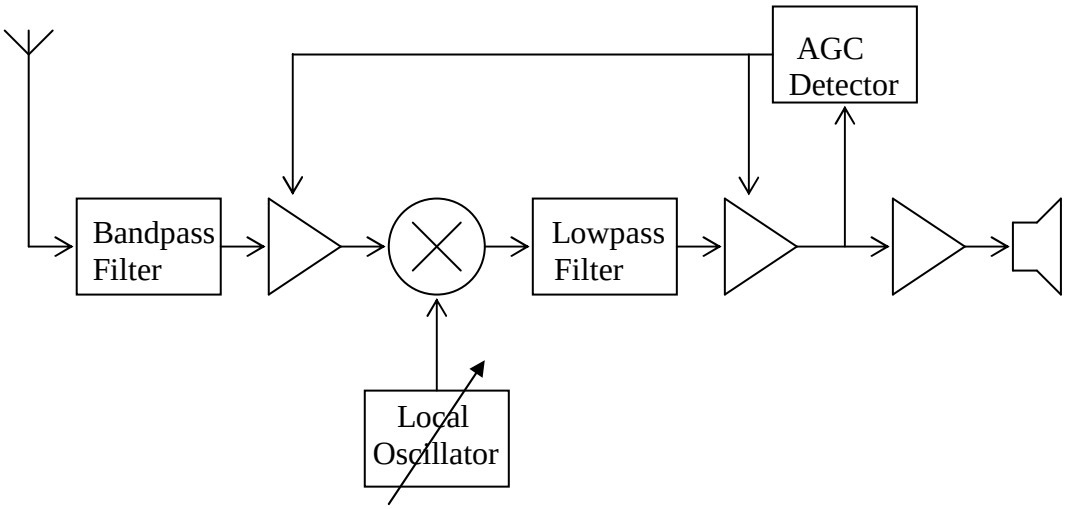
Many TRF receivers use *regeneration*, which means feeding some of the signal from the output of the RF amplifier back to its input, in such a way as to reinforce the signal at the input of the RF amplifier. This technique is a form of *positive feedback*. It has the benefit of increasing the amplification of the RF amplifier (because some of the signal "circulates" through it many times, being amplified each time) and also increasing the selectivity, since the signal also passes through the bandpass filter at the output of the RF amplifier many times. Of course an amplifier with positive feedback is an oscillator, so if too much regeneration is applied, the circuit will oscillate. Regenerative receivers (a name for TRF receivers that use regeneration) usually have a control to adjust the amount of regeneration, which is adjusted to get the maximum possible sensitivity and selectivity without oscillation.

The advantage of TRF receivers is that they are simple to construct and require relatively few components—typically just two or three transistors and a handful of other parts. This simplicity made them attractive in the days before transistors, when thermionic tubes were used for amplification in radio receivers, as tubes were relatively expensive, so the fewer the better!

Their big disadvantage is that they have very poor selectivity and dynamic range. Tunable bandpass filters just aren't capable of rejecting an unwanted signal that is only a couple of kilohertz away from the signal you are listening to, so unwanted signals will also get through to the detector and be recovered as audio or cause inter-modulation distortion. TRF receivers are also best suited for receiving AM signals. Although regenerative receivers can be used with CW and SSB signals, by adjusting the regeneration control so the circuit just barely oscillates, adjustment is tricky and the quality of reception poor. For these reasons, TRF receivers are no longer widely used.

23.4 The Direct-Conversion Receiver

A design that is used in quite a few homebuilt receivers is the Direct Conversion (DC) receiver. In a DC receiver, the radio-frequency signal from the antenna is mixed with a locally generated oscillator signal, producing the usual sum and difference mixing products. The frequency of the oscillator that generates this local mixing signal—it is known as the *local oscillator* (LO) or *beat frequency oscillator* (BFO)—is set so the difference mixing product is at audio frequency. In this way, the DC receiver "directly converts" the desired radio-frequency signal to audio, where it can be filtered and amplified. Let's look at the circuit in a little more detail.

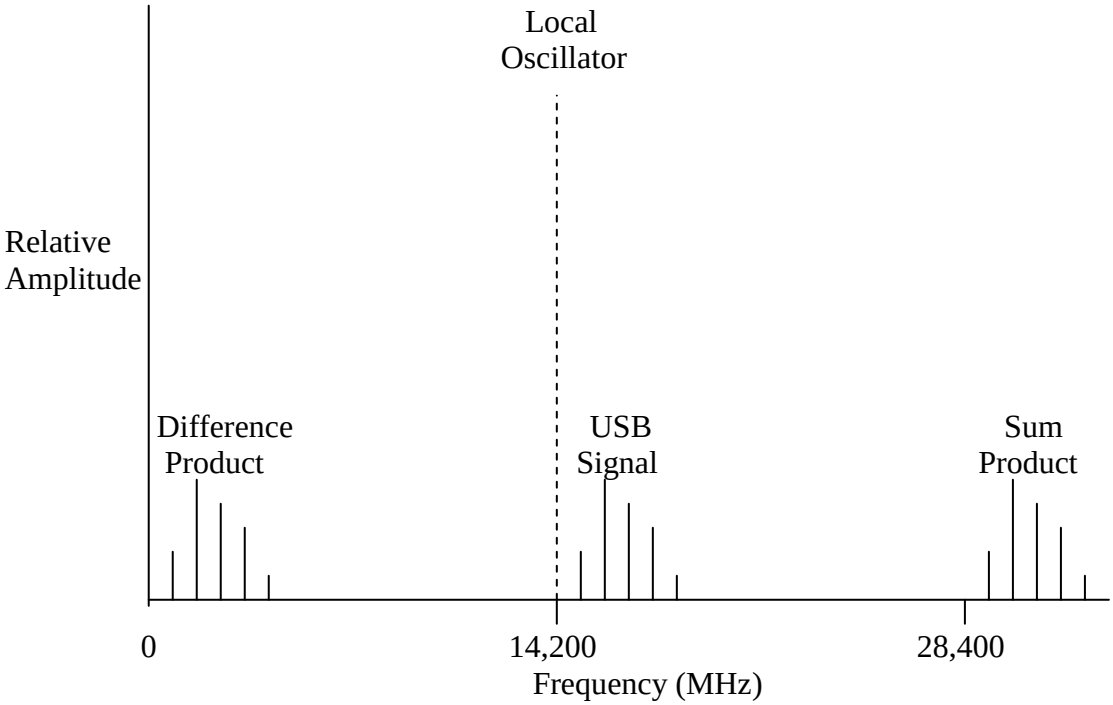


A Direct-Conversion Receiver

The signal from the antenna first passes through a bandpass filter. Unlike in the Tuned Radio Frequency receiver, this bandpass filter is not responsible for the overall selectivity of the receiver. Its role is simply to reject interference from strong local commercial broadcast stations and the like. It does not have to be tunable—usually a fixed-tuned filter covering an entire amateur band will suffice.

The signal is then amplified by an RF amplifier and fed into the product detector, which is represented on the diagram using the symbol for a mixer—the circle with a cross in it. “Mixer”, “Modulator” and “Product Detector” are different names for essentially the same circuit, depending on the exact role it plays. The product detector mixes the amplified RF signal with a signal generated by the tunable local oscillator, generating the usual sum and difference mixing product.

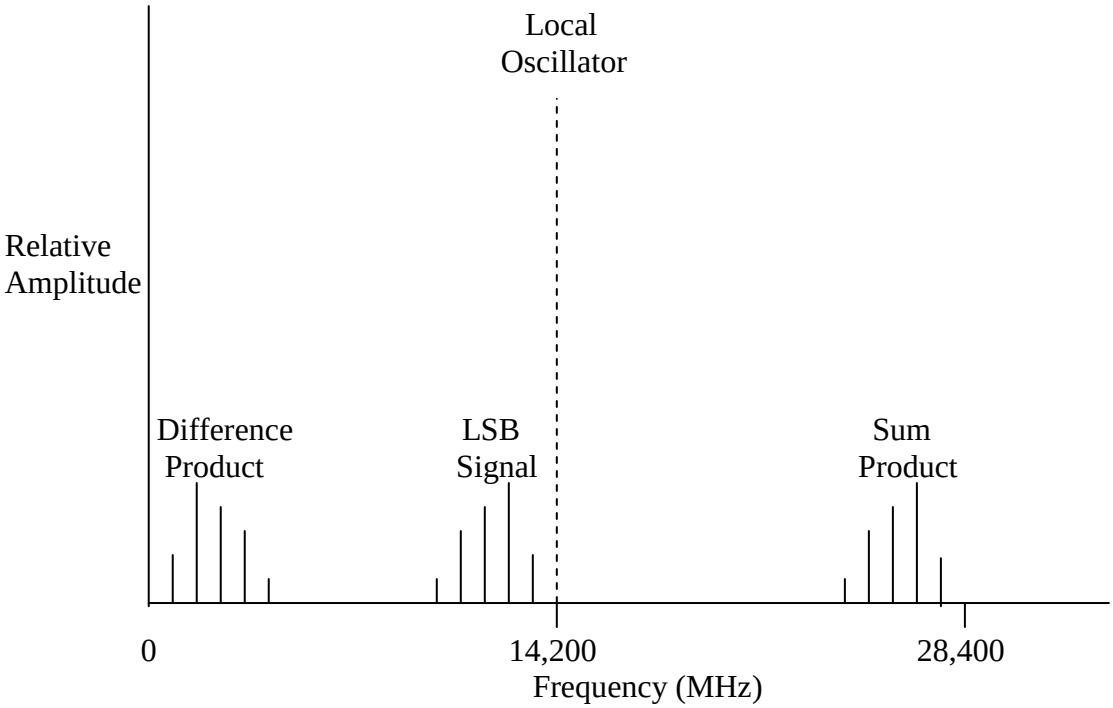
Suppose we want to receive an upper sideband signal on 14,200 MHz. By convention, we refer to the frequency of a single-sideband signal as the frequency where the carrier would have been if it had not been suppressed. This frequency is known as the *pseudo-carrier* frequency. The upper sideband of this USB signal will span a frequency range from 14,200 3 MHz to 14,203 0 MHz, or 300 Hz to 3 kHz above the pseudo-carrier frequency. If the local oscillator is set to exactly 14,200 MHz, the difference mixing products will range in frequency between 300 Hz and 3 kHz. What we have done is to translate the USB signal from its frequency of 14,200 MHz back to the baseband.



Receive mixer output for a USB signal

This graph shows how mixing the 14,200 MHz USB signal with a 14,200 MHz signal from the local oscillator generated a difference mixing product (signal frequency – local oscillator frequency) in the audio range and a sum product (signal frequency + local oscillator frequency) up above 28,400 MHz.

Although the example used a USB signal, the same process would work equally well using an LSB signal, and the local oscillator frequency would still be 14,200 MHz, the pseudo-carrier frequency. The following graph shows the same process with a lower sideband signal.



Receive mixer output for an LSB signal

Once again the difference product is back in the audio frequency range, while the sum product is at around twice the signal frequency, at 28,400 MHz. Also note how for the lower sideband signal, the mixing process has inverted the sideband (so the recovered audio is the mirror image of the sideband), which makes up for the sideband inversion that would have occurred when the LSB signal was generated.

So whether the signal is USB or LSB, mixing it with a local oscillator at the pseudo-carrier frequency will demodulate it and recover the audio to the baseband.

To complete the hat trick, suppose we have a CW signal at 14,200 MHz. All we need to do is set the local oscillator just below it—say at 14,199 4 MHz, which is 600 Hz below the CW signal—and the difference mixing product will be a 600 Hz tone, just right for listening to CW. So we can also use the product detector to receive a CW signal. Setting the local oscillator 600 Hz above the CW signal would work just as well.

We now pass the recovered audio through a lowpass filter. The main purpose of the filter is to remove the difference mixing product from signals near to the one that we are listening to. For example, suppose there is a CW signal at 14,205 MHz while we are listening to our 14,200 MHz USB signal. The difference mixing product of the 14,205 MHz CW signal and the 14,200 MHz local oscillator is 5 kHz—in other words, we have translated the unwanted CW signal downwards in frequency to the audio range just as we have translated the wanted USB signal to audio. However, a lowpass filter with a cutoff frequency of around 3 kHz or so should be able to remove the unwanted CW signal without affecting the desired USB signal.

Because it is quite easy for a strong signal to overload a mixer, causing inter-modulation distortion, the gain ahead of the mixer (i.e. the gain of the RF amplifier) is usually kept low so as not to amplify unwanted strong signals and overload the mixer. Most of the gain in a Direct Conversion receiver is at audio frequencies, in the amplifiers following the lowpass filter.

The only remaining part of the circuit is the Automatic Gain Control (AGC) system. Because there is such a wide range of signal strengths on the air, and because the strength of a particular signal can vary with propagation changes, it is useful to have some way of automatically controlling the gain of the receiver. There must be a lot of gain to amplify weak signals, which must be reduced to avoid overloading when strong signals are present. While this effect could be achieved with a manually operated gain control, but the poor operator would have to constantly work at it, and may occasionally be punished by a painfully loud signal when the gain is not reduced quickly enough. Also, when tuning from a strong signal (with the gain turned right down) to a weak signal, a weak signal can be missed altogether unless the gain is turned up first.

The solution is AGC. The AGC detector samples the audio signal after the first audio amplifier, and automatically adjusts the gain of the RF amplifier and the audio amplifier to keep the output signal level fairly constant. The output signal is then amplified by a final audio power amplifier and used to drive headphones or a speaker. The AGC control voltage is often also used to drive a *signal strength meter*, known as an "S meter", that indicates the strength of the received signal using a fairly arbitrary scale calibrated from S1 (a very weak signal) to S9 (a very strong signal). S meters are generally not well calibrated, but S9 is often taken as being 100 µV, with every S unit below that level representing about 6 dB.

The DC receiver has several advantages over a TRF receiver. Most importantly, its selectivity is very good, because unwanted nearby signals are easily filtered out by the audio lowpass filter that follows the product detector. It is more stable, having no tendency to oscillate like regenerative TRF receivers do. And it is easy to receive single sideband and CW signals with a DC receiver—you just tune the signal in, without having to fiddle with the regeneration control.

However, the DC receiver does have one significant disadvantage. Since the same local oscillator frequency can be used to tune either a upper sideband or a lower sideband signal, if you are listening to say an upper sideband signal and there is a different signal occupying the frequencies on the other side of the local oscillator where the lower sideband would have been, the other signal will also be shifted to audio frequencies and will interfere with the station you are trying to listen to.

For example, suppose you are listening to an USB signal at 14,200 MHz as before, but there is also a CW signal at a frequency of 14,199 MHz. Mixing the 14,200 MHz local oscillator signal with the 14,199 MHz CW signal results in a 1 kHz audio tone. Since this tone falls within the same 300 Hz to 3 kHz audio range as the desired USB signal, you cannot filter it out using the lowpass filter. And because the unwanted signal is so close in frequency to the desired signal, you can't use the RF bandpass filter to reject it either.

The unwanted signal on the other side of the local oscillator signal is called an "image", so the principal disadvantage of the Direct Conversion receiver can be described as its inability to reject images, or lack of "*image rejection*". There are more sophisticated variations of the basic Direct Conversion design that *are* able to reject images, but these are quite complex and fall outside the scope of this course.

Finally, a DC receiver must be designed carefully to ensure that the LO is isolated from the antenna port. Some simple DC receivers radiate some of the LO through the antenna, causing interference to listeners near the reception frequency.

Summary

The key attributes of a receiver are sensitivity, selectivity and dynamic range. Sensitivity is the ability to receive weak signals; selectivity is the ability to distinguish between adjacent signals; and dynamic range is the ability to receive weak signals despite the presence of strong signals nearby.

In the tuned radio frequency (TRF) receiver all signal filtering is done at radio frequencies. As a result they have poor selectivity. Regeneration, which consists of feeding some of the output signal back to the input of the RF amplifier, can increase both the sensitivity and selectivity of the TRF receiver, but makes it prone to oscillation. The oscillation, if well-controlled, can be used to facilitate CW and SSB reception.

In the direct-conversion (DC) receiver, the incoming RF signal is mixed down to audio frequency using a product detector and local oscillator. Most of the selectivity of a DC receiver is contributed by audio filters following the product detector. DC receivers have much better selectivity than TRF receivers, but they suffer from an image response to the opposite sideband that can only be eliminated with complex designs. A bad DC design may also radiate some of the local oscillator, causing interference to other users.

Signal to noise ratio (SNR) determines whether a signal is readable or not. Noise can originate within the receiver or on the band. The receiver has a *noise figure* (in dB), which can also be expressed as a *noise temperature* (in K). At HF and below, band noise normally limits the SNR. At VHF and above, receiver noise is normally the limiting factor. Special semiconductors, feedlines and techniques are required to minimise receiver noise at these frequencies.

Revision Questions

1 The specification "1 µV to provide better than 20 dB signal to noise ratio in a passband of less than 1 kHz" refers to:

a. Sensitivity.
b. Selectivity.
c. Stability.
d. Image rejection.

2 The ability of a receiver to extract weak signals and amplify them to a readable level is known as the receivers':

a. Sensitivity.
b. Selectivity.
c. Q factor.
d. Gain factor.

3 The sensitivity of a communications receiver can best be varied by:

a. Altering the input voltage.
b. Altering the RF gain.
c. Changing the IF.
d. Adjusting the volume control.

4 The dynamic range of a receiver can be described as:

a. Its audio output.
b. The tuning range.
c. The operating voltage.
d. The range of signal strengths over which it operates satisfactorily.

5 The RF stage of a receiver is used to:

a. Improve its sensitivity.
b. Improve its selectivity.
c. Change the frequency.
d. Change the signal tone.

6 The ability of a receiver to receive the desired signal whilst rejecting other frequencies is known as:

a. Sensitivity.
b. Selectivity.
c. A tuning scale.
d. Wavelength.

7 The circuit that lowers a radio receivers' gain as the received signal becomes stronger is known as:

a. AGC.
b. Filter.
c. ALC.
d. Selector.

8 What is an S-meter?

a. A meter used to measure sideband suppression.
b. A meter used to measure spurious emissions from a transmitter.
c. A meter used to measure relative signal strength in a receiver.
d. A meter used to measure solar flux.

9 The output from a direct conversion receiver is the difference in frequency between:

a. The BFO and the incoming signal.
b. The BFO and the local oscillator.
c. The mixer and IF frequencies.
d. The incoming signal and the local oscillator.

10 A radio receiver that amplifies and filters the incoming signal at RF and then uses a diode detector to demodulate an AM signal would be called:

a. A superhet receiver
b. A crystal set
c. A tuned radio frequency receiver
d. A direct-conversion receiver

11 A preamp with a low noise temperature is advertised for use with HF receivers. You are tempted to buy one.

a. Go ahead and buy it—it will markedly improve your weak-signal reception through the static on the low bands.
b. Go ahead and buy it—it will markedly improve your weak-signal reception on the 10 m band.
c. Go ahead and buy it—it will markedly improve your general DX performance
d. Save the money—at HF, it will not improve receiving performance at all.

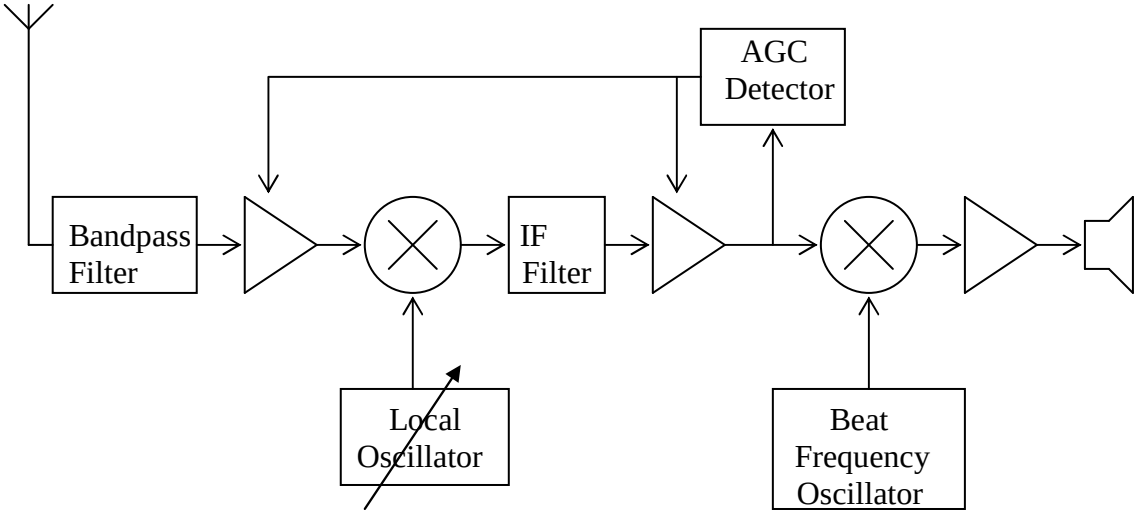
12 You are shopping for a receiver for weak-signal UHF applications. You should buy the one that features:

a. A high noise temperature.
b. A high noise figure.
c. High dynamic range.
d. A low noise temperature.

Chapter 24: The Superheterodyne Receiver

24.1 The Single-Conversion Superhet

The superheterodyne receiver or "Superhet" as it is commonly known has the most widely used receiver design in amateur radio[6]. It overcomes the lack of image rejection of the DC receiver by converting the incoming RF signal to one or more *intermediate frequencies* before demodulating it. The block diagram of a typical *single-conversion superhet* (one with only a single intermediate frequency) is shown below.



A Single-Conversion Superhet Receiver

The RF signal from the antenna is first filtered by a bandpass filter. As in the DC receiver, this filter can be a fixed-tuned filter covering an entire amateur band, since the receiver does not rely on this filter (known as the *preselector*) for its selectivity. As we shall see, the main purpose of the preselector is to reject the image frequency. The signal is then amplified in an RF amplifier – once again, not too much amplification, to avoid overloading the mixer that follows. In some designs the RF amplifier may be omitted entirely.

In the first mixer, the RF signal is mixed with the signal from the tunable local oscillator. However, instead of mixing the modulated signal down to baseband, it produces an *intermediate frequency* (IF). Common intermediate frequencies for single-conversion superhets are 455 kHz, 9 MHz and 10,7 MHz. The modern trend is to go for higher IFs, as IF stage performance improves.

Suppose we want to receive a signal on 14,200 MHz again, and the intermediate frequency is 9 MHz. We could use a local oscillator frequency of either 5,200 MHz (because the difference between 5,200 MHz and 14,200 MHz produces the IF of 9 MHz) or 23,200 MHz (because the difference between 14,200 MHz and 23,200 MHz is also the IF of 9 MHz). For this example, we will assume that we chose a local oscillator frequency of 5,200 MHz, since this frequency is within the range that can easily be generated by a VFO.

The resulting 9 MHz IF signal is then filtered by the IF filter, which is a very narrowband bandpass filter. Modern designs typically use crystal filters, so for this example we shall assume a crystal filter with a passband of 9,000 3 MHz (300 Hz above 9 MHz) to 9,003 0 MHz (3 kHz above 9 MHz). Signals within the passband will be passed with little

[6] This dominance is likely to be challenged by DSP (digital) receivers in the near future. However, the DSP algorithms use superhet principles to achieve the same results.

attenuation, while signals that fall outside the passband will be blocked. So what components of our original RF signal will fall within the filter passband? An RF signal at 14,200 3 MHz would be mixed down to 9,000 3 MHz by the 5,2 MHz local oscillator signal; and a signal at 14,203 0 MHz would be mixed down to 9,003 0 MHz. So the signals that originated at these frequencies—from 14,200 3 to 14,203 MHz—will make it through the IF filter. This range corresponds to the USB signal at 14,200 MHz.

What about signals on the "other side" of 14,200 MHz, from 14,197 0 to 14,199 7 MHz, i.e. the frequencies that would have caused an image in a Direct Conversion receiver? Well, they will be mixed down to between 8,997 0 MHz and 8,999 7 MHz, and will be rejected by the IF filter, so they do not cause a problem.

There is still an image, but in this case it is from 3,800 3 MHz to 3,803 0 MHz. A 3,800 3 MHz signal mixed with our 5,2 MHz local oscillator will generate an additive (sum) product at 9,000 3 MHz, and a 3,803 0 MHz signal will generate a mixing product at 9,003 0 MHz. So signals within the frequency range 3,800 3 MHz to 3,803 0 MHz when combined with the 5,2 MHz local oscillator signal will also generate products in the IF range from 9,000 3 to 9,003 0 MHz that will be passed by our IF filter. However, this time the image is far away from the desired signal at 14,200 MHz, so it can easily be filtered out before the mixer. This filtering is the main purpose of the preselector. It must pass the desired frequencies, around 14,2 MHz, while rejecting the image frequencies, around 3,8 MHz. Fortunately, because these frequencies are so far apart, it is fairly easy to get good image rejection from a simple passive bandpass filter made only of inductors and capacitors.

Note that by varying the frequency of the local oscillator we can change the RF frequency that will be mixed down to the 9 MHz IF. For example, a local oscillator frequency of 5,3 MHz would mix an RF signal of 14,300 MHz down to the 9 MHz IF, while our original reception frequency of 14,200 MHz would now be mixed down to 8,900 MHz and would be blocked by the IF filter. You tune a superhet receiver by varying the frequency of its local oscillator, just like for a DC receiver.

The circuitry after the IF filter is virtually identical to that of a DC receiver. The IF signal is amplified, and then mixed with another locally generated oscillator signal—this time called the "Beat Frequency Oscillator" or BFO—to recover the audio signal, which is then amplified by an audio amplifier. Since the IF signal is at a fixed frequency—e.g. 9 MHz—the BFO does not have to be tunable so we can use a stable fixed-frequency 9 MHz crystal oscillator for the BFO.

The AGC also works similarly to that of a direct conversion receiver, although in this case the AGC control voltage is derived from the intermediate frequency, rather than the audio frequency output. This AGC is IF-derived, audio-derived AGC in the DC receiver. IF-derived AGC is superior to audio-derived AGC as it is able to respond more rapidly to sudden changes in signal strength.

The same design can be used to receive CW signals as well. For example, to receive a CW signal with a frequency of 14 200 MHz, the local oscillator would be set to 5 199 4 MHz, generating an IF signal at the difference between these frequencies, 9,000 6 MHz, which is within the passband of the crystal filter. After being amplified it will be mixed with the 9 000 MHz BFO signal in the product detector, generating an audio tone of 600 Hz.

So how about LSB signals? The simplest approach would be to have a second IF filter with a passband from 8,997 0 MHz (3 kHz below 9 MHz) to 8,997 7 MHz (300 Hz below 9 MHz) that can be selected in place of the 9,000 3 to 9,003 0 MHz filter when we want to receive an LSB signal. Then when switching from USB to LSB, all you have to do is switch

filters, the local oscillator and BFO frequencies remain the same. Since crystal filters are expensive, an alternative approach is to use the same IF filter for LSB and USB reception, and just change the frequencies of the local oscillator and BFO. For example, to receive a LSB signal at 14,200 MHz using the 9,000 3 to 9,003 MHz IF filter, we could set the local oscillator to 5,196 7 MHz and the BFO to 9,003 3 MHz. The reader can fill in the details as an exercise.

Since we can receive USB, LSB and CW signals using this design, how about AM signals? There are two options. The simplest is just to leave the receiver design exactly as it is, and receive AM signals as though they were single-sideband signals, ignoring the carrier and the other sideband, which will be filtered out by the IF filter. A better approach would be to provide another selectable IF filter, this time with a passband from 8,997 to 9,003 MHz to accommodate the 6 kHz bandwidth of an AM signal. The product detector would then be designed so that in the absence of any signal from the BFO, it would act as a half-wave rectifier and would detect AM by rectifying the IF signal (an "envelope detector"). We now have the benefits of "proper" AM demodulation, including accurate reproduction of the frequencies of the original audio signal even if the receiver is not accurately tuned.

24.2 Multiple-Conversion Superhet Receivers

When choosing the IF for a single-conversion superhet, there is a trade-off between image rejection and selectivity. It is easier to make highly selective filters at a low IF—say 455 kHz. However, a low IF means that the image frequency is close to the desired frequency, making it difficult to effectively reject the image with a simple preselector. Conversely, a high IF produces a large separation between the image frequency and the desired signal, making it easy to reject the image while passing the desired signal. However, a high IF makes it harder to achieve the desired selectivity.

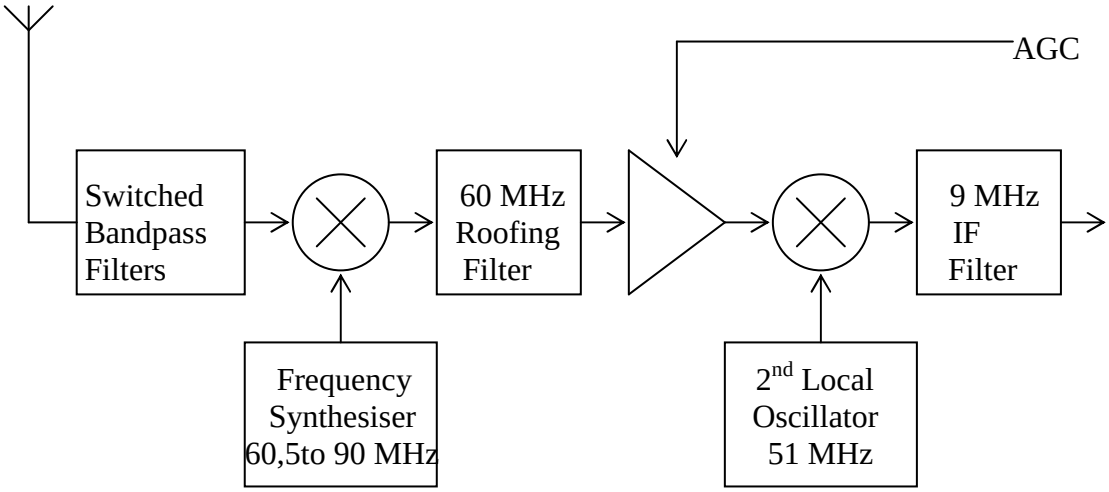
The classical solution to this dilemma has been to use a superhet design with *two* intermediate frequencies—a high *first IF* for good image rejection, followed by a low *second IF* for good selectivity. However, modern crystal filters generally make this unnecessary in HF receivers, since very good selectivity is available from crystal filters at intermediate frequencies in the 9 MHz region, which is a high enough IF to attain good image rejection as well. Of course in VHF and UHF receivers, a higher first IF may be required to prevent unwanted image responses.

Nevertheless, the multiple-conversion superhet is still the most common approach for commercial HF receivers, but for a slightly different reason. Most commercial receivers and transceivers today offer "general coverage receive", meaning that they can receive on any frequency in the MF and HF bands, typically from 500 kHz to 30 MHz, or even up into the UHF range (up to 1 GHz). Unfortunately, this versatility gives them a problem with IF leak-through, which occurs when the first mixer is not exactly balanced, allowing some of the original RF signal to appear at the IF output. If the RF signal is at the same frequency as the IF, it will be passed by the IF filter, causing the radio to respond to a frequency that it shouldn't, a phenomenon known as a "spurious response". This response would not be a big problem for an amateur-bands-only receiver, because an IF like 8,5 MHz could be chosen well away from any amateur band. Then the preselector, possibly assisted by a dedicated notch filter at the IF, will be able to reject incoming RF signals at the IF, so there are no signals in the RF input that could "leak through" into the IF stages.

However, the designer of a general-coverage receiver is not so fortunate. If the chosen IF is anywhere in the receiver's frequency range, it will be impossible to reject RF signals at the IF, since these might include the frequency the receiver is tuned to! The solution is to choose an IF that is either above or below the receiver's frequency range. However, now the selectivity versus image rejection tradeoff comes back with a vengeance because a filter that is above the frequency range of a typical general coverage HF receiver—that is, above

30 MHz—will not have the necessary selectivity; while a filter at an IF that is below the receiver's coverage—say 455 kHz—will not allow adequate image rejection.

The usual solution is a multiple-conversion superhet where the first IF is *above* the receiver coverage range, allowing good image rejection and IF leak-through rejection, while the second IF is at a lower frequency where better selectivity can be obtained. This arrangement is known as an *up-conversion* design, since the incoming signal is first converted up to a higher frequency. The IF filter at the high first IF is often referred to as a *roofing filter* and is generally wide enough to permit signals of all modes through, up to 12 or 15 kHz in the case of a receiver that supports FM as well as other modes. Much narrower filters are provided for the different modes (e.g. a 6 kHz filter for AM and a 2,4 kHz filter for SSB) at the lower second IF. The block diagram below shows the "front end" (the circuitry from the antenna to the IF filter) of a typical general-coverage dual-conversion superhet.



Front-End of a General Coverage Dual-Conversion Superhet

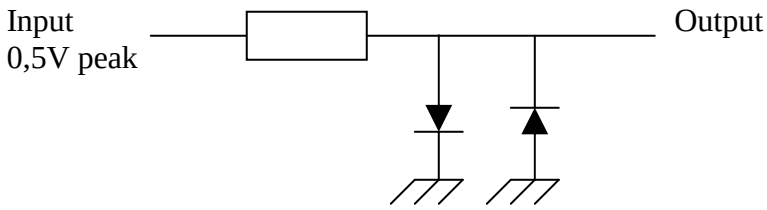
The design includes a bank of switched bandpass filters in the preselector, to allow coverage of the range 0,5 to 30 MHz with good image and IF leak-through rejection. The first local oscillator is a frequency synthesiser running from 60,5 to 90 MHz, which up-converts the RF signal to the first IF of 60 MHz. Here it is filtered by the roofing filter, which would typically have a bandwidth of 12 kHz or so. The purpose of the roofing filter is to reject signals which are close enough to the desired frequency to be passed by the preselector, but which might cause either inter-modulation distortion or an image response in the second mixer. The IF signal is then amplified and converted back down to the second IF frequency of 9 MHz. From here on the circuitry would be similar to the single-conversion design featured earlier.

As the term "multiple-conversion superhet" suggests, there is no reason why more than two IFs could not be employed, resulting in triple-conversion and quadruple-conversion superhets.

24.3 Noise Limiters and Noise Blankers

Many common sources of amplitude-modulated noise generate amplitude "spikes" of short duration but high amplitude, which extend over a wide range of frequencies. These may contain substantial energy due to their large amplitude, even though their duration is short. Such noise is generated both by natural sources, such as thunderstorms, and by man-made ones, like inadequately suppressed ignition systems. Interference from these noise sources can be reduced by noise limiters and noise blankers, which are available on almost all modern amateur transceivers.

A noise limiter is a very simple circuit that limits the maximum amplitude of the received signal.



Noise Limiter

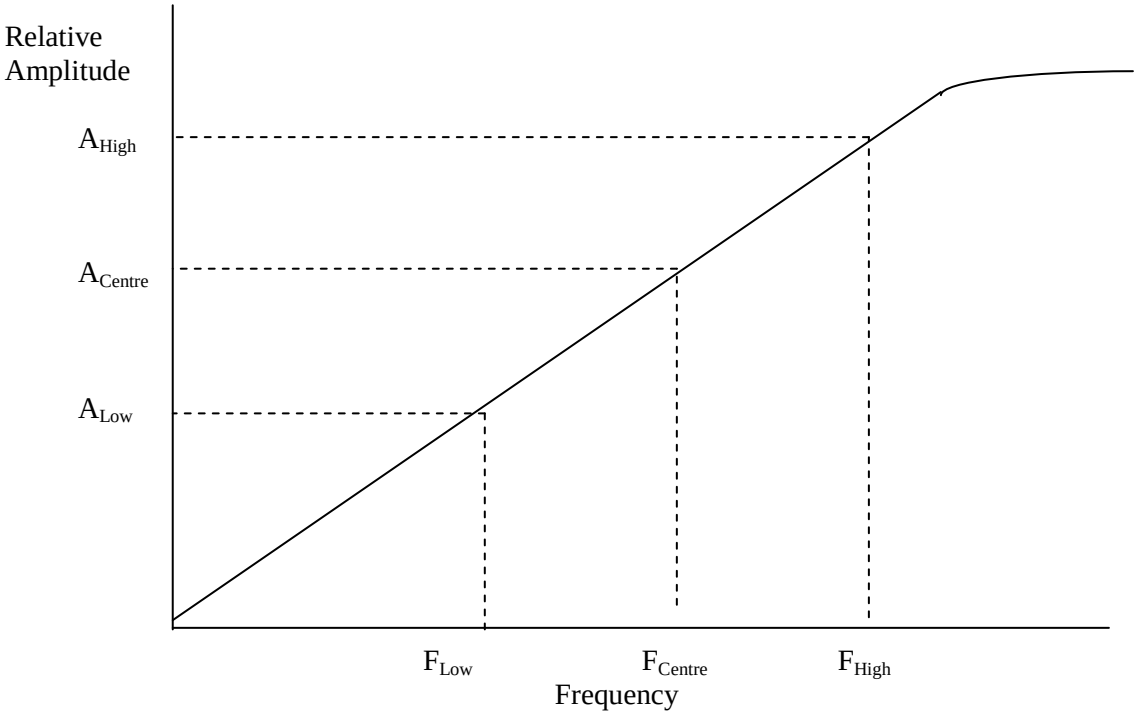
Assume the input signal has a maximum amplitude of 500 mV peak under normal circumstances. This voltage is less than the 600 mV forward bias voltage of the diodes, so they do not conduct, and the input signal will be passed to the output unchanged. Suppose a noise pulse generates a signal amplitude of 5 V. As soon as the amplitude exceeds 600 mV, the diodes conduct, effectively limiting the maximum output to 600 mV peak and substantially reducing the energy of the noise signal.

The noise blanker is a more sophisticated variation on this idea. It detects the large amplitude of the incoming noise signal, and then immediately mutes (turns off) the audio output of the receiver completely for a predetermined time, typically a few milliseconds. Although the desired signal is blocked along with the noise, the human ear is quite insensitive to very short gaps in sounds, and the resulting signal degradation is much less than would have been caused by the high amplitude noise spike.

24.4 Frequency Modulation (FM) Reception

The basic superhet design can also be used to receive frequency modulated (FM) signals. However, in this case, the product detector is replaced by a *Foster-Seeley discriminator* or a *ratio detector*. These are circuits that convert frequency variations into a varying output voltage, so recovering the modulation from an FM signal.

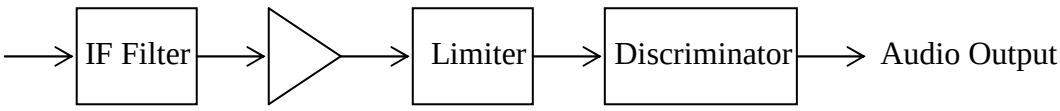
The discriminator works by positioning the FM signal on the slope of a selective filter, so that variations in the frequency of the FM signal will result in variations in its amplitude. This converts the FM into a combined AM and FM, and a simple diode detector is used to recover the modulation from the AM component.



How the slope of a filter can be used for FM detection

The graph shows how the slope of a highpass filter could be used to convert FM into AM. As the signal frequency increases from F_{Centre}, the centre frequency, to $F_{High,}$ the amplitude of the output increases from A_{Centre} to A_{High}. If the frequency decreases from F_{Centre} to F_{Low}, the amplitude of the output ill also decrease, from A_{Centre} to A_{Low}.

Because the discriminator is also sensitive to changes in the amplitude of the incoming signal, it should be preceded by a *limiter*. The limiter is a circuit that limits the amplitude of the signal, so that amplitude variations are not passed on to the discriminator or ratio detector that follows. The limiter circuit is identical to the noise limiter discussed earlier, except that in an FM receiver the circuit would be driven at a much higher input level, causing the diodes to conduct and clamp the output signal to 600 mV peak. In this way the output of the limiter will always be at the same level (600 mV peak), irrespective of the amplitude of the input signal. The block diagram below shows the final IF stage of a typical FM receiver



Final IF Stage of an FM Receiver

When the received signal is very weak, the limiter is ineffective and the discriminator will respond to amplitude variations, which cause hiss in the audio. As the signal gets stronger and the limiter takes effect, the hiss decreases, a process called “quieting”. In order to prevent the hiss from bothering the listener when there is no received signal, most FM receivers incorporate a squelch feature, which mutes (turns off) the audio output when the received signal is below a minimum level known as the squelch threshold. The squelch threshold may be fixed or it may be adjustable using a squelch control.

24.5 Reciprocal Mixing

Superhet receivers typically do a good job in rejecting unwanted signals. Unfortunately, there is one exception.

In the section on synthesisers, it was mentioned that synthesisers suffer from some phase jitter, which creates crud around the intended output frequency. Instead of there being only a single line on the spectrum display, there is now a line with some surrounding crud. This crud mixes with unwanted signals and effectively spreads them out across the entire band. The phenomenon is known as reciprocal mixing. The receiver performance is measured by either measuring the rise in noise level when an interfering signal is nearby (but out of the passband), or by having a wanted signal in the passband, then displacing it by a known frequency increment and increasing its amplitude until it presents the same amount of power in the passband. Reciprocal mixing is a practical measure of dynamic range, and is at least 100 dB for most good-quality receivers.

Summary

The superhet receiver converts the incoming RF signal to one or more *intermediate frequencies* before demodulating it. Superhet receivers have an *image frequency* that when mixed with the local oscillator will also generate the same IF as the desired receive signal. The image frequency will be either the sum of, or the difference between, twice the IF frequency and the desired receive frequency. The role of the preselector is to reject incoming RF signals at the image frequency, preventing them from causing a spurious (unwanted) response in the receiver. The choice of intermediate frequency is a tradeoff between selectivity (better at low LF) and image rejection (better at high IF). If a single IF cannot give adequate selectivity and image rejection, a dual conversion design may be employed, with a higher first IF to give good image rejection, and a lower second IF to give good selectivity.

Noise limiters limit the amplitude of pulse noise, reducing the effect on the receiver. Noise blankers mute the audio output for a short time (a few milliseconds) when the higher amplitude associated with pulse noise is detected.

FM signals are detected using a *Foster-Seeley discriminator* or *ratio detector*. The discriminator should be preceded by a limiter to prevent it from being affected by variations in the amplitude of the signal. Weak FM signals have a characteristic hiss on them, and as the signal strength increases and the limiter becomes effective the hiss goes away, a process known as *quieting*. Most FM receivers incorporate a *squelch* function, which mutes the audio output when there is no received signal to avoid the annoying hiss.

Synthesiser phase noise manifests itself as a spreading out of unwanted signals. Some of the crud ends up in the receiver passband, causing wanted signals to be less readable. The result is reduced dynamic range, as a wanted weak signal cannot coexist with nearby loud signals.

Revision Questions

1 In an FM receiver, the effect of sufficient signal arriving to start the limiter operating, thus reducing background noise, is known as:

a. Damping.
b. Squelch.
c. De-emphasis.
d. Quieting.

2 The selectivity of a receiver is mostly controlled by:

a. Gain of IF and RF stages.
b. Bandwidth of RF and IF stages.
c. Sensitivity of RF and IF stages.
d. Stability of RF and IF stages.

3 In a superheterodyne receiver intended for AM reception, what stage combines the received radio frequencies with energy from a local oscillator to produce an output at the receiver's intermediate frequency?

a. The mixer.
b. The detector.
c. The RF amplifier.
d. The AF amplifier.

4 In superheterodyne receivers, the setting of the first IF is governed by two general principles:

a. High IF gives good image rejection but low IF gives better selectivity.
b. High IF gives good image rejection and good selectivity.
c. Low IF gives good image rejection and high IF gives good selectivity.
d. Low IF gives good image rejection and good selectivity.

5 The function of an IF amplifier in a superheterodyne receiver is to:

a. Improve its sensitivity.
b. Improve its selectivity.
c. Buffer the mixer output.
d. Amplify the loudspeaker output.

6 How can the selectivity of an IF amplifier be improved?

a. Varying the supply voltage.
b. Varying its resistance.
c. By use of a bandpass filter.
d. By use of a lowpass filter

7 The detection of a Single Sideband signal in a receiver requires a:

a. Carrier Insertion Oscillator.
b. Special Aerial.
c. An SSB amplifier.
d. A special transformer.

8 A superheterodyne receiver is operating with its local oscillator on the high side of the incoming signal. If its IF is 450 kHz and it is receiving an input signal of 14 100 kHz, an image will be produced:

a. At this tuning point if a strong signal is on 15 MHz.
b. Further up the tuning band if a strong signal is on 15 MHz.
c. At this point if a strong signal is on 13 650 kHz is present.
d. Further up the tuning band if a strong signal is on 13 650 kHz.

9 In a single conversation superheterodyne receiver with an IF of 450 kHz, a signal is received first at 12 000 kHz and then again at 12 900 kHz. These two received signals are called:

a. Cross-modulation products.
b. Band spread products.
c. Image signals.
d. De-emphasis signals.

10 The process by which a receivers' local oscillator and mixer resonant circuits maintain a constant IF separation is known as:

a. Tracking.
b. Isolation.
c. Shielding.
d. Attenuation.

11 The preferred circuit for resolving an SSB signal is

a. A product detector.
b. A fullwave rectifier.
c. A Colpitts oscillator.
d. A crystal oscillator.

12 The product detector is used to

a. Detect square waves.
b. Balance out noise signals.
c. Deduce unwanted feedback.
d. Resolve SSB and CW modulation.

13 What is the purpose of the detector in a receiver?

a. To amplify the incoming signal.
b. To operate the squelch circuit.
c. To operate the on/off light.
d. To demodulate the modulating signal.

14 Electrical interference on reception can best be limited by means of a:

a. Squelch circuit.
b. Noise Limiter.
c. Isolation transformer.
d. Decoupled loudspeaker.

15 The receive facility that switches off the audio circuit in the absence of satisfactory signal strength is:

a. A noise limiter.
b. A squelch circuit.
c. A VOX circuit.
d. An AF gain control.

16 In order to avoid image reception on VHF receivers they normally have:

a. Low IF frequencies.
b. Crystal controlled local oscillators.
c. A stable BFO.
d. High IF frequencies.

17 A dual-conversion superhet receiver contains:

a. Two IF amplifiers of different frequencies.
b. Two RF pre-amplifiers.
c. Stereo audio circuits.
d. Two antenna connections.

18 You are listening to a signal, which is comfortably readable. Suddenly, the passband fills up with noise, making the signal unreadable. You tune around and find a very loud local station about 20 kHz away. The station is known to be clean, and his transmitter is probably not at fault. The problem with your receiver is probably:

a. Poor selectivity.
b. Poor sensitivity.
c. Poor dynamic range.
d. Poor image rejection.

Chapter 25: Transceivers and Transverters

25.1 The Transceiver

Although in the early days of amateur radio transmitters and receivers were usually separate, in modern practice these are usually combined in a single piece of equipment, the *transceiver*. Transceivers have several advantages over separate transmitters and receivers:

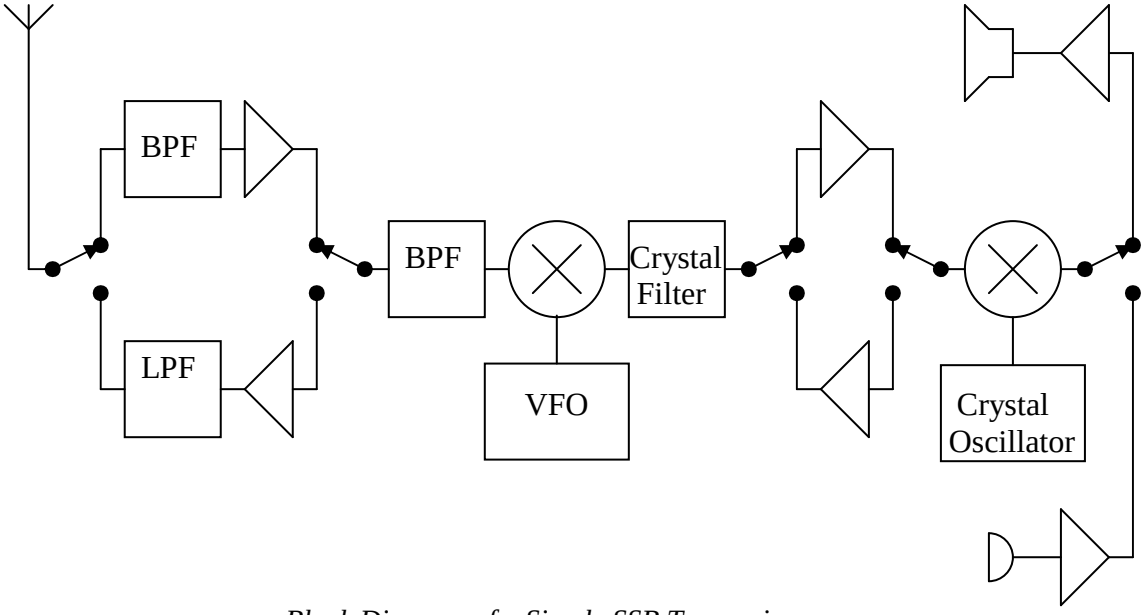
- It is easier to make the transmitter and receiver tune together, so the user does not have to separately tune the transmitter and receiver to the same frequency.

- Much of the circuitry including the oscillators or synthesisers, filters, antenna switching, microprocessor and display can be shared between the transmitter and receiver, making transceivers less expensive than a separate transmitter and receiver.

- Installation is simpler, with less space and fewer cables required.

For these reasons, almost every modern amateur transmitter also includes receive capability, at least on the bands that it can transmit on. Many transceivers also offer "general coverage" receive, being able to receive signals outside the amateur bands.

In order to maximise the sharing of components between the transmitter and receiver section of a transceiver, they will typically use the same frequency conversion scheme but in reverse. For example, if the receiver is a double-conversion superhet with IFs at 60 MHz and 9 MHz, the transmitter will probably generate SSB using the filter method at 9 MHz (allowing reuse of the 9 MHz crystal filter), and then mix it up to 60 MHz using the receiver's BFO, and then mix it back down to the actual transmit frequency (using the same synthesiser for both transmit and receive). Circuit reuse is further enhanced since popular balanced mixers (such as the passive diode mixers) are essentially reversible – a signal injected at the RF port will mix with the local oscillator to generate a signal at the IF port, while a signal injected at the IF port will mix with the local oscillator to generate a signal at the RF port. Many filters will also work equally well in either direction. The diagram below shows a simple SSB transceiver that reuses several of its functional blocks. "LPF" stands for "Lowpass Filter", "BPF" for "Bandpass Filter" and "VFO" for "Variable Frequency Oscillator".



Block Diagram of a Simple SSB Transceiver

There are two different signal paths through the transceiver, depending on the position of the transmit/receive switches. In practice, the switches would probably consist of relays or solid state switches. With the switches in the position shown, the transceiver is in receiving mode. The signal from the antenna is filtered by a bandpass filter, amplified in the RF amplifier, fed through another bandpass filter and then mixed down to IF by the signal from the VFO. This IF passes through a narrow crystal filter which removes all frequencies other than the desired one. The filtered signal is then amplified in the RF amplifier and finally demodulated in the product detector, using the signal from the crystal oscillator that serves as a beat frequency oscillator (BFO) for the receiver.

On transmit (with the switches all switched the other way) the signal from the microphone is amplified by the preamplifier, and then fed into the detector, which this time serves as a balanced modulator. The output of the balanced modulator is amplified, filtered using the crystal filter to remove the unwanted sideband, and mixed to the final output frequency using the signal from the VFO. The output is passed through a bandpass filter to remove the unwanted mixing product, then amplified by the RF power amplifier and finally filtered to remove any harmonics. This design reuses the mixer, product detector, crystal filter, VFO, crystal oscillator and one of the bandpass filters.

Most amateur transceivers are designed to operate into an unbalanced antenna with an impedance of 50 Ω.

25.2 The Transverter

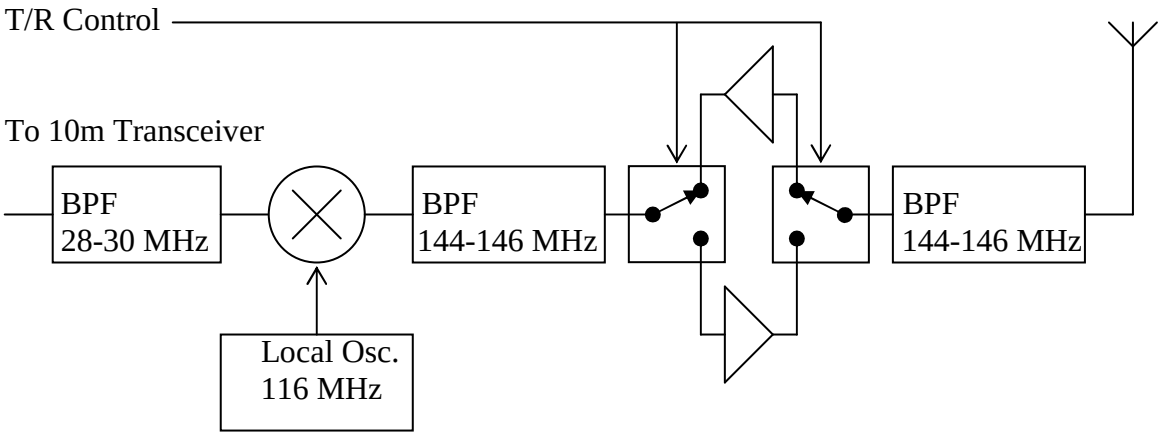
A *transverter* consists of a receive converter and a transmit converter, normally packaged in a single enclosure. The transverter allows a transceiver to transmit and receive on a frequency band other than the ones it was originally designed for. It does so by incorporating a LO and mixer that can convert the output of the transceiver to the new frequency band, and convert signals received on this band to a frequency that the transceiver can receive on.

An amateur might own a good HF transceiver, but wants to be active on the 2 m band. A 2 m transverter can convert frequencies in the range 28 to 30 MHz (around the 10 m amateur band) to the frequency range 146 to 148 MHz (the 2 m band). It could do so by mixing the signals with a 116 MHz local oscillator signal. On transmit, the sum of the local

oscillator and a low-level transmit signal from the transceiver in the range from 28-30 MHz would give an output in the range 144 to 146 MHz, while on receive the difference between an input signal in the range 144 to 146 MHz and the 116 MHz local oscillator signal would provide an output in the range 28 to 30 MHz. The operator would have to remember that a displayed frequency of 28,xxx MHz corresponds to an actual operating frequency of 144,xxx MHz, while 29,yyy corresponds to 145,yyy.

Most transverters are designed to operate with low power on transmit, generally around 0 dBm (1 mW), and include a power amplifier to amplify the signal. Some transceivers have a special connector for transverters, which provides a low-level RF signal to drive them. Otherwise an attenuator should be used on transmit to decrease the transceiver's output to a level that the transverter can safely handle. Most transverters will be damaged by the typical 100 W output of most HF transceivers.

Like transceivers, transverters usually use components like the local oscillator, mixer and filters for both transmit and receive. Their transmit/receive switching is usually controlled by the transceiver, using its T/R control output. A block diagram of a typical 2 m transverter is shown below. It includes both a power amplifier for transmit, and a receive preamplifier to amplify weak signals and compensate for losses in the mixer.



Block Diagram of a 2 m Transverter

On transmit, the low-level signal from the transceiver is filtered by the 28 to 30 MHz bandpass filter, and then mixed with a 116 MHz local oscillator. The mixer generates a "sum" product in the range 144 to 146 MHz, and a "difference" product in the range 90 to 88 MHz (inverted). The bandpass filter that follows the mixer rejects the difference product. The 144 to 146 MHz product is amplified by a power amplifier, and passed through a final bandpass filter to remove any harmonics.

On receive, the signal from the antenna is filtered by the 146 to 148 MHz bandpass filter, amplified by a low-noise preamplifier, and filtered again to remove any image signals in the 88 to 90 MHz range. It is then mixed with the 116 MHz local oscillator, generating a "sum" product at 262 to 264 MHz, and a "difference" product in the range 28 to 30 MHz. The 28 to 30 MHz bandpass filter rejects the unwanted "sum" product, leaving only the 28 to 30 MHz signal that is fed to the transceiver.

Some transceivers allow the frequency display to be offset when using a transverter. In our previous example, the transceiver display could read the actual operating frequency of 144 to 146 MHz while using the transverter, while the transceiver was actually being tuned from 28 to 30 MHz.

One final comment: A receive converter can be seen simply as an extra stage to a multiple-conversion superhet, turning the receiver's front end into a variable-frequency first IF, the receiver's first IF into a second IF and so on. Likewise, the transmit converter simply provides another stage of up-conversion to the total transmitter.

Summary

A transceiver consists of a transmitter and receiver for combined into one. They are widely used because of operator convenience and the lower cost that can be achieved by using the same components for both transmit and receive functions.

Transverters convert a transceiver to transmit and receive in a new frequency band. They work by mixing the output of the transceiver or the input from the antenna with a fixed frequency local oscillator, translating the frequency to the new band. Transverters generally require an input signal power of about 1 mW when transmitting, and care should be taken not to overdrive them.

Revision Questions

1 The advantage of a transceiver is that it:

a. Costs less than a separate transmitter and receiver.
b. Better integrates frequency control of the transmitter and receiver.
c. Is more compact and easier to install than a separate transmitter and receiver.
d. All of the above.

2 A transverter provides the ability to:

a. Convert a transformer to more than one voltage.
b. Change the units to measure a transmitter's power output.
c. Use an existing transceiver to cover another amateur band.
d. Receive extra-terrestrial signals.

3 A receive converter converts a single-conversion superhet into:

a. A TRF receiver.
b. A DC receiver.
c. A dual-conversion superhet.
d. A triple-conversion superhet.

Chapter 26: Antennas

26.1 Antennas and Electromagnetic Fields

Antennas convert electrical energy—which generally requires conductors to carry it—into electromagnetic energy, which is able to radiate through space. They can also convert the electromagnetic energy back into electricity, for further processing in a receiver.

An electrical current flowing in a conductor generates a magnetic field in the space around the conductor. This effect is the principle behind electromagnets. If the current flowing in the conductor varies with time, the magnetic field around the conductor also varies with time. However, according to the principle of induction, a varying magnetic field gives rise to an electric field; and conversely, a varying electric field gives rise to a magnetic field. So by varying the current in a conductor, we can create a varying magnetic field, which will in turn create a varying electric field, which will create a varying magnetic field, and so on. The resulting interrelated varying electric and magnetic fields are called electromagnetic waves, and can travel long distances. At the frequencies that we are interested in, these electromagnetic waves are radio waves, although heat, light, x-rays and gamma rays are also examples of electromagnetic waves of higher frequencies.

In electromagnetic waves, the electric and magnetic fields are perpendicular (at right angles) to each other, and both are perpendicular to the direction of motion of the wave. For a radio wave travelling horizontally, if the electric field is horizontal with respect to the surface of the earth, the magnetic field will be vertical; while if the electric field is vertical, the magnetic field will be horizontal. We refer to the *polarisation* of electromagnetic waves according to the orientation of the electric field. If the electric field is horizontally oriented, the wave is horizontally polarised; while if the electric field is vertically oriented, the wave is vertically polarised. In physics, the electric field is referred to as the E-field, and the magnetic field as the H-field.

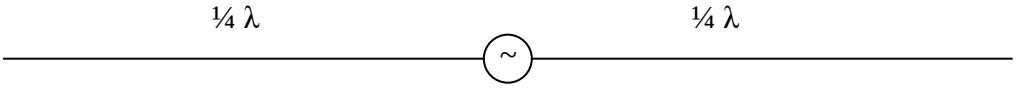
The orientation of the electric field (and hence the polarisation of the radio waves) generally corresponds to the orientation of the conductor carrying the current that generated the radio waves. Generally, an antenna consisting of horizontal conductors will generate horizontally polarised radio waves, while an antenna consisting of vertical conductors will generate vertically polarised radio waves.

Polarisation is important because a vertically polarised antenna will not respond to horizontally polarised radio waves, and *vice versa*. For line-of-sight communications, it is important that the polarisation of the transmitting antenna should be the same as that of the receiving antenna.

When radio waves travel through space, they carry energy. To recover this energy, an antenna must capture as large an area of the wavefront as possible. Every antenna has a *capture area*, which is roughly the surface area from which the antenna can capture energy. The more effective the antenna is (as described later in the section on gain), the larger its capture area is.

26.2 The Half-Wave Dipole

The half-wave dipole is a simple antenna that consists of a half wavelength of wire, fed in the centre by a radio-frequency voltage source.



A Half-Wave Dipole

Assume that an alternating voltage is applied at the dipole's resonant frequency – that is, the frequency for which each side of the dipole is exactly a quarter wavelength (¼λ).

Consider an instant in time when one side of the voltage source is positive and the other side is negative. The effect will be to pull some electrons out of the side of the dipole that is attached to the positive terminal of the voltage source, and push some electrons into the side that is attached to the negative terminal of the voltage source. Since electrons repel each other, as you force electrons into the wire at the negative terminal of the voltage source, they will repel the electrons that are already in the wire, pushing them towards the end of the wire. As each electron moves up a bit, it will repel its neighbour, forcing it to move up too, and so on causing a wave to travel down the wire. This wave travels at the speed of light until it reaches the end of the wire, where the electrons cannot bunch up any more because they have nowhere to go. At this point, the wave will reflect from the end of the wire and head back towards the feed-point.

The effect is similar to what you would observe if you set up a line of pool balls with one end against the cushion, and then knocked the one furthest from the cushion into its neighbour. Each ball hits the next one, and so on down the line, until the last ball, which is up against the cushion so it cannot move. The "pool-ball wave" is then reflected from the cushion and travels back in the opposite direction. Of course the pool-ball wave does not travel at the speed of light!

Similarly in the dipole, applying an instantaneous potential difference at the feed-point (the place where the voltage source is connected to the antenna) causes waves to travel from the feed-point to both ends of the wire, where they are reflected and head back towards the feed-point. On one side—the side where the negative potential is applied—the wave consists of an increase in the electron density, "compressed electrons" (like the pool-ball analogy). On the other side—the side where the positive potential is applied—the wave consists of a reduced electron density as electrons have been attracted out of the wire by the positive potential.

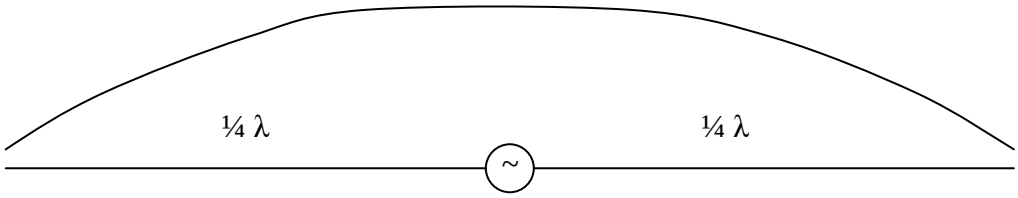
These waves both have to travel ¼λ to the end of the wire, and another ¼λ in the opposite direction before they get back to the feed-point, a total distance of ½λ. One half-cycle later, the waves will return to the feed-point, heading in the opposite direction. But because it is half a cycle later, the voltage source will have the opposite polarity, so it will now be pushing the electrons in the opposite direction. But this is the same direction that the reflected wave is traveling in, so the reversed polarity of the voltage source will reinforce the reflected wave. Another half cycle later the waves will have reflected off the ends of the wire again, and the voltage source will again have reversed polarity, so once again the voltage source will reinforce the waves of increased and reduced electron density that are coursing up and down the wire like water sloshing about in a bath.

Because the applied voltage is always reinforcing the waves, a fairly small voltage can (over a few cycles) cause a large movement of electrons, in other words a large current (since the electrons are charge carriers that are flowing backwards and forwards, and so

constitute an alternating current flowing in the antenna). Remembering that from Ohm's Law $R = V/I$, we see that the feed-point resistance will be low, since a small V causes a large I.

The fact that the resistance is not zero means that the antenna is dissipating power. Where is this power going? It is being radiated as radio waves from the antenna. This apparent resistance of the antenna caused by energy being radiated from it is called the *radiation resistance* of the antenna. The radiation resistance of a half-wave dipole is about 72 Ω.

You will also note that in the centre of the antenna, the electrons are quite free to move, so a large current will flow. However, the nearer you get to the ends of the antenna, the less free the electrons are to move, up to the points right at the ends, where the electrons can hardly move at all. This means that there will be a larger current flowing in the centre of the antenna than at the ends. If you superimpose a graph of the amplitude of the current flowing on top of a diagram of the antenna, you get something like this:



The Current Distribution in a Half-Wave Dipole

As you can see, the current is strongest in the centre of the dipole and tapers off towards the ends. Because it is the current flowing in the antenna that is primarily responsible for the emission of radio waves, these waves can be visualised as being emitted from the point of highest current at the centre of the antenna. This fact has some practical implications. For example, the ends of a dipole can be bent back without affecting its properties as an antenna much, since the ends are relatively unimportant as far as radiation is concerned.

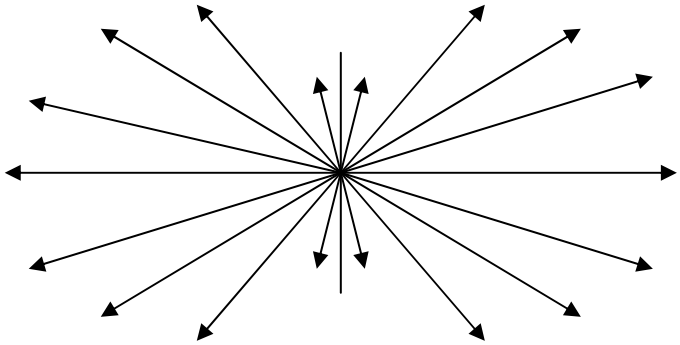
Although the current in a dipole decreases towards the ends, the opposite happens with the voltage—it is greatest at either end of the dipole. In general, points of high current (current *peaks*) are points of low voltage (voltage *nodes*), and points of low current (current *nodes*) are points of high voltage (voltage *peaks*). So beware the open ends of antennas, where no current is flowing, as they invariably carry high voltages!

This current distribution—and the corresponding voltage distribution—are called "standing waves" since they have a wave-like shape (roughly like a sine wave) but the points of highest and lowest current and voltages do not move—they are standing still. Another way to think of the standing waves is that they are caused by the interaction between the waves from the feed-point moving towards the ends of the antennas, and the reflected waves moving back towards the feed-point.

A final point about half-wave dipoles: Sometimes it is more convenient to have the feedpoint off-centre. The reason may be mechanical, because the centre of the dipole is hard to reach, or electrical, because a more convenient feed impedance is desired. As can be seen from the current and voltage distributions, the impedance will be much higher than in the middle. Little difference should result in the antenna's performance, provided that the energy can be efficiently coupled into the dipole with a suitable feedline.

Our next task is to calculate in what directions the dipole will radiate, or the *radiation pattern* of the antenna. This task is quite easy for an antenna that has only a single point of maximum current, like the dipole. When an alternating current flows in a wire, it radiates

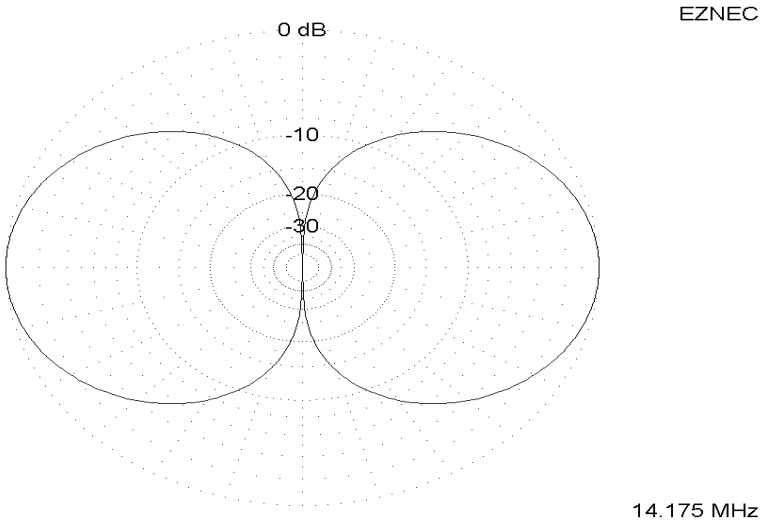
most strongly at right angles to the wire, and less strongly the further you move from the right-angled direction. This situation is depicted below, where the length of each arrow represents the strength of the radiation coming from the wire in every direction.



Radiation Pattern of a Dipole

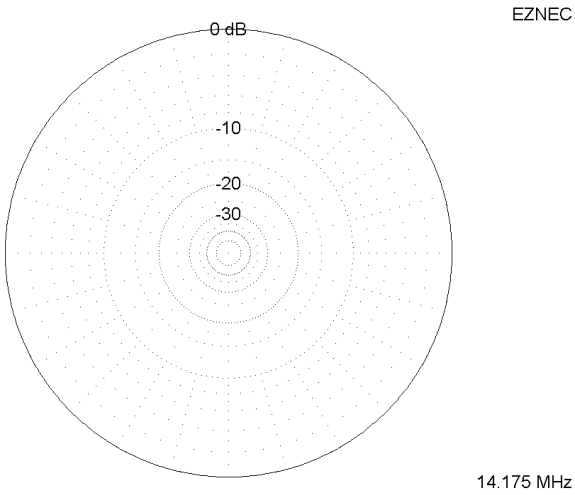
In this case, the dipole is the vertical line in the centre of the diagram. Note how the strongest radiation is perpendicular (at right angles) to the wire, and the strength of the radiation decreases as the angle gets further from perpendicular. There is no radiation at all along the axis of the wire (i.e. vertically up or down the page in this diagram).

Actually we don't usually draw radiation patterns with arrows like this. Instead we just draw a line that would join the tips of all the arrowheads. In other words, the distance from the centre of the diagram to the line indicates the strength of radiation in that direction. The following diagram is the radiation pattern of a dipole, drawn in the conventional way.



Radiation Pattern of a Dipole in Free Space

Although the dipole is not shown, it is oriented the same as in the previous diagram—vertically on the page in the centre of the plot. The plot lines indicate the relative strength of the field in each direction: The further the distance from the centre of the diagram to the line, the stronger the radiation in that direction. Note the nulls (points of minimum radiation) off the ends of the dipole (vertically up and down in the plot). Of course, the dipole will radiate equally in all directions perpendicular to it, not just in the two directions (left and right) shown in the diagram. So if you think of it in three dimensions, it would look like a doughnut shape with the wire in the middle. We can also look at the dipole end-on. Seen end on, the pattern is like this:

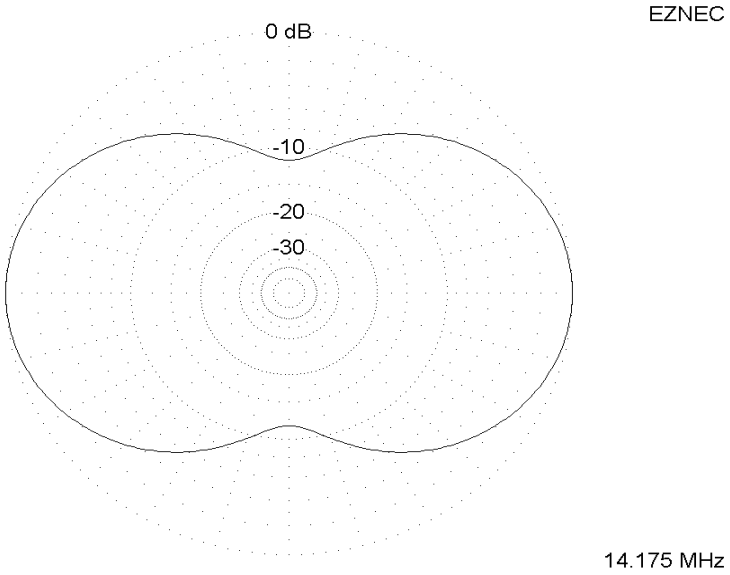


Radiation pattern of a Dipole in free space viewed end-on

Here the dot in the middle represents the wire seen end-on, and the circular radiation pattern indicates that it radiates equally in all dimensions.

These diagrams show the radiation pattern of a dipole in *free space*, i.e. far away from the ground. The ground reflects radio waves, so in order to understand the radiation pattern of antennas mounted over ground (like all normal amateur antennas), we also need to take into account the effect of these reflections. In general, the waves reaching a distant point will come from two sources—a direct wave from the antenna, and a wave that has been reflected by the ground. Depending on the difference between the distances traveled by these waves, they may reinforce each other, or they may cancel each other out.

Assume that we orient the dipole horizontally and mount it some distance above ground. Then the pattern viewed from above, would look pretty much the same as the free-space pattern.

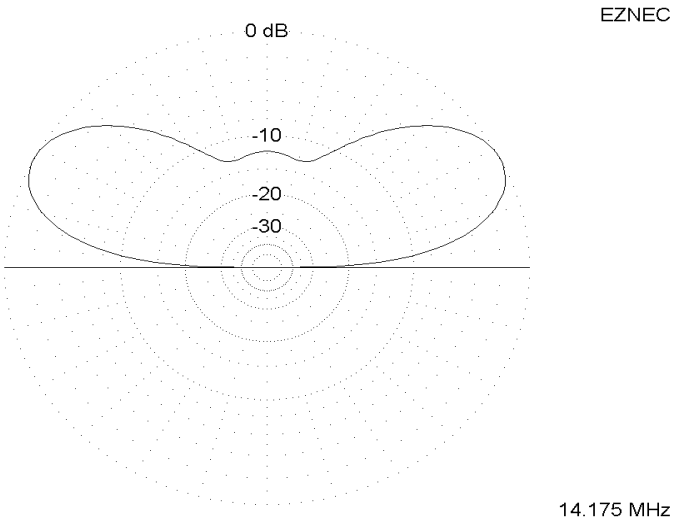


Azimuth of a horizontal half-wave dipole near ground, viewed from above

Once again the dipole is oriented vertically on the page, in the middle of the diagram, although this time we are looking down on it from above. This is what we call an “azimuth pattern”. The main effect of the ground reflections has been to “fill in” the nulls (directions of zero radiation) off the ends of the dipole, although there is still much less radiation from

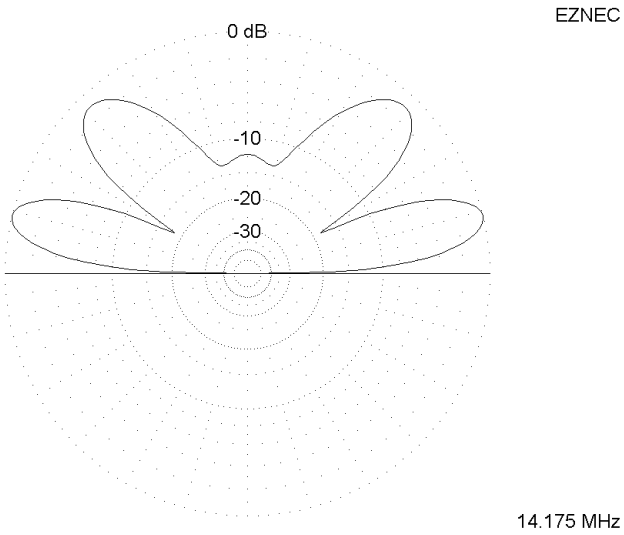
the ends of the wire than perpendicular to it—approximately 13 dB less according to the scale on the diagram. For this reason, we talk about the horizontal dipole as being a *bi-directional* antenna, favouring two directions (left and right in the diagram above) at the expense of the others.

The nearby earth has a much greater effect on the vertical pattern of the antenna. Instead of radiating equally in all directions, as it would with no ground present, we get the following pattern:



Elevation pattern of a horizontal half-wave dipole near ground, viewed end-on

The ground reflections have cancelled out most of the low-angle and high-angle radiation, leaving one "lobe" on each side, with an angle of maximum radiation of about 27º in this instance. The exact pattern depends on the height of the antenna above ground. In this case, it is ½λ wavelength above ground. If we raise it to 1λ, the elevation pattern now looks as follows:



Elevation pattern of a horizontal half-wave dipole 1λ above ground, viewed end-on

There are now two lobes in each direction, one at a fairly low angle (14º) and one at a higher angle (47º). In general, the higher we raise the dipole above ground, the more lobes we get, and the lower the elevation angle of the lowest-angle lobe.

Low-angle radiation is very desirable for making long-distance contacts. If making long-distance contacts is a priority, and you are using horizontally-polarised antennas like this dipole, the rule is simple: the higher, the better.

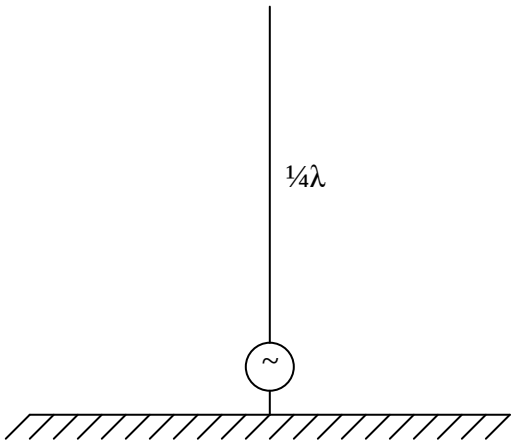
The dipole is normally used as a horizontal antenna, but it can also be constructed vertically if desired, especially in the VHF and UHF bands where the wavelength is shorter and the height required more reasonable.

In practice, because the tips of a dipole carry little current and contribute little to the total radiation, a dipole can be deformed to fit into confined spaces. Tips can be folded at right angles or even doubled back on themselves. If only a single support is available, the feedpoint can be attached to the support and the tips sloped towards the ground. This configuration is called an "*Inverted V*", for obvious reasons. Provided that the angle is not too acute, the Inverted V functions almost like a classic dipole.

A horizontal half-wave dipole is sometimes called a "Hertz" antenna, since this type of antenna was first used by Heinrich Hertz[7], the German physicist who also gave his name to the unit of frequency.

26.3 The Quarter-Wavelength Vertical

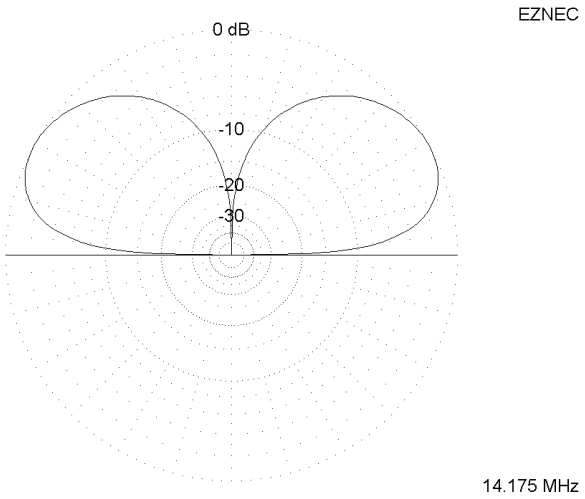
Another popular antenna is the quarter-wavelength vertical. It consists of a quarter wavelength of wire mounted vertically and driven at its base, with other side of the driving voltage connected to ground. It works like "one half of a dipole", with the return path for the current being through the ground.



A Quarter-Wavelength Vertical

There's not much point in talking about the radiation pattern of a ¼λ vertical in free space, since it needs the ground as one of its connections! The elevation plot of the quarter-wave vertical over ground is shown below.

[7] Not to be confused with Love Hertz, the 1975 hit song by Nazareth and Jim Capaldi.



Elevation pattern of a quarter-wavelength vertical over real ground

The azimuth pattern (the pattern when viewed from above) would just be a circle, since the antenna radiates equally in all directions[8] when viewed from above. Antennas that radiate equally in all (horizontal) directions are called *omni-directional* antennas.

Although simple in theory, the quarter-wave vertical has a difficult practical problem to overcome: It is difficult to create low-impedance ground connections at radio frequencies. A ground rod of the kind used for household mains grounds will generally not present a very low impedance at radio frequencies, and the resistance of the resulting ground connection will cause power to be dissipated as heat in the ground rather than being radiated from the antenna. Although a quarter-wave vertical will work with just a ground-rod, you could easily find 75% or more of the power applied to the antenna being wasted heating the ground instead of being radiated.

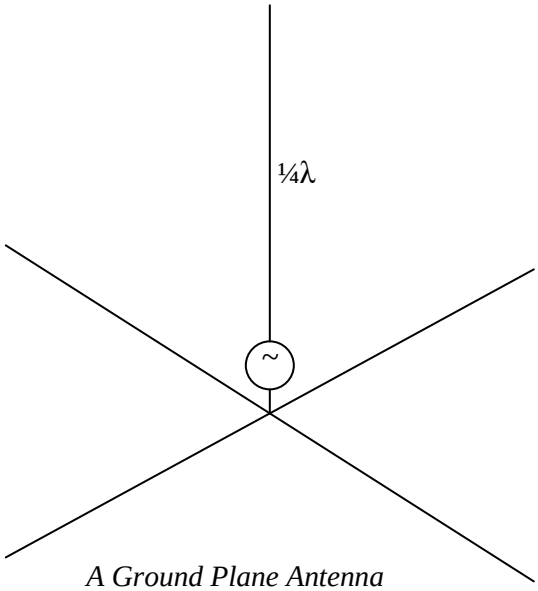
The radiation resistance of a ¼λ vertical antenna is 36 Ω, half of the value for a half-wave dipole. You can understand why; electrically, the antenna looks exactly like a halfwave dipole if you look at it from above the groundplane, as the groundplane will reflect the image of an out-of-phase ¼λ element. However, the effective ground resistance will also contribute to the impedance seen at the feed-point, which will usually be quite a bit higher than 36 Ω.

A ¼λ vertical is sometimes called a “Marconi” antenna, since this type of antenna was first used by Guglielmo Marconi, the inventor of wireless telegraphy and the pioneer of transatlantic radio communications.

26.4 The Ground Plane Antenna

One solution to the problem of creating a low-impedance RF ground connection is to raise the vertical antenna 1/8 λ or more above the ground, and feed it against some radials, rather than against ground. The radials effectively act as the “missing side” of the dipole, but because they run in opposite directions, the radiation from them cancels, leaving only the radiation from the vertical wire, which is called the *radiator*.

[8] A snide remark you will sometimes hear is that it radiates equally badly in all directions…



A Ground Plane Antenna

For ease of feeding a ground-plane antenna, ¼λ resonant radials are generally used. Three or four radials suffice for simple installations. The radials may be laid out flat (as shown in the diagram) or they can droop downwards. With flat radials, the radiation resistance will be between 20 Ω and 26 Ω depending how high above ground the antenna is mounted. Drooping the radials will increase the feed-point impedance, so with the right angle of droop (about 45º) the impedance can be raised to 50 Ω, which is a good match for the coax cables usually used to feed such antennas. Of course, if you droop the radials to 90°, you will have made yourself a dipole, with a radiation resistance of 72 Ω!

The radiation pattern of a "ground-plane" antenna is identical to that of a quarter-wave vertical.

In general, the more radials you add, the longer they are and the better they are insulated from the ground, the more efficient the groundplane antenna becomes. In AM broadcasting, most regulators require 120 radials buried ¼λ radials, based on research conducted by RCA in the 1930s. More recent research has indicated that better results can be obtained with insulated radials, and that there is an optimal relationship between radial length and the number of radials required.

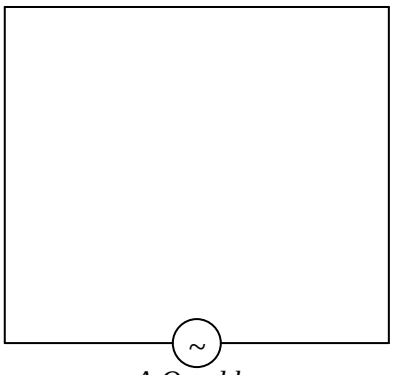
26.5 Short Antennas

An antenna that is shorter than the length normally required for resonance will have some capacitive reactance. Both a dipole shorter than ½λ and a vertical antenna shorter than ¼λ will exhibit some capacitive reactance. This reactance can be cancelled out by an equivalent amount of inductive reactance in series with the antenna. This inductance may be provided by an inductor known as a *loading coil*, which may be placed at the base of the antenna (*base loading*) or in the centre of the element (*centre loading*). Centre loading is more efficient electrically than base loading, but makes the mechanical design of the antenna more difficult as it has to support the weight of the loading coil.

Similarly, antennas that are slightly longer than required for resonance will have some inductive reactance, which can be tuned out using a series capacitor, but this is less common. This technique can be useful when matching a groundplane antenna to a 50 Ω cable. By making the vertical element longer than ¼λ and using some series capacitance, a good match to 50 Ω can be obtained.

26.6 Loop Antennas

A full-wavelength loop is another simple antenna. Loops come in all shapes: squares, rectangles and triangles to name a few. The diagram below shows a “quad loop”, which has a square shape, with each side being ¼λ.

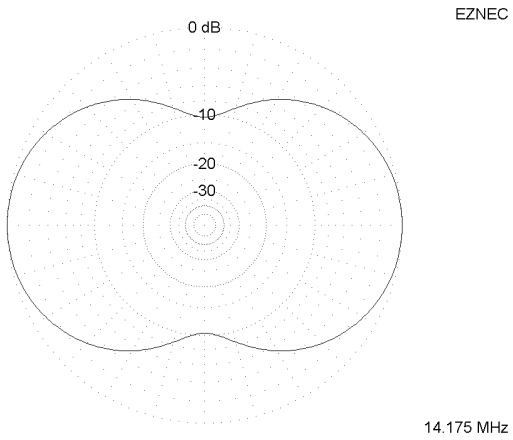


A Quad loop

Each side of the loop is ¼λ, so the total length of the loop is 1λ. In loops, there is no end of the antenna for the wave to reflect from, so it just keeps on going around the loop until it arrives back at the feed-point still traveling in the same direction. If the length of the loop is one wavelength, the applied voltage will once again be in the same direction, so the wave traveling around the loop is reinforced, giving it a relatively low and purely resistive feed-point impedance. The radiation resistance of a loop is a bit more than that of a dipole, since there is more wire to radiate electromagnetic waves, typically about 130 Ω.

The loop is normally erected vertically. There are two points of high current in the loop – one at the feed-point, and one halfway around the loop. If the feed-point of the loop is in the middle of one of the horizontal wires, as shown in the diagram, both points of high current will be carrying horizontal currents and the resulting radiation will be horizontally polarised. If on the other hand the loop was fed in the middle of one of the vertical wires, the points of maximum current would be carrying currents flowing in a vertical direction and the resulting radiation would be vertically polarised. Feeding the loop at one of the corners results in diagonal polarisation.

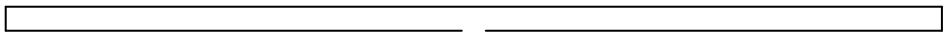
The radiation pattern of a horizontally polarised one-wavelength loop is similar to that of a dipole, with the strongest radiation being perpendicular to the horizontal wires of the loop. This result is to be expected since the loop is effectively two horizontal dipoles vertically spaced by ¼λ, with their ends bent out of shape.



Azimuth pattern of a horizontally polarised quad loop

26.7 Folded Dipole

The folded dipole is a one-wavelength loop that has been "flattened" into a shape similar to a dipole.



Feed point

A folded dipole

Its radiation pattern is the same as that of a dipole, but the radiation resistance of a folded dipole is about 300Ω. It is higher than a normal dipole because of the transformation effect of the wire adjacent to the basic dipole.

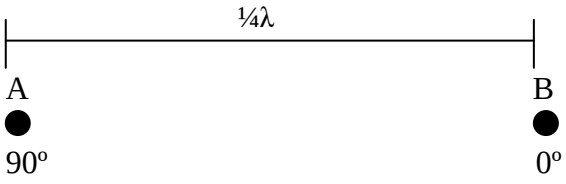
26.8 Multi-element arrays

So far, the antennas we have considered have all had a single radiating element. These antennas are practical and very simple to build, but have limited *directivity*—that is, the ability to favour one direction at the expense of the others. Single-element vertical antennas are *omni-directional*, radiating equally in all directions, while the dipole and quad loop are *bi-directional*, favouring two directions at the expense of the others.

However, if we know the location of the radio station we want to contact, and are able to point our antenna in that direction, it would be better to have an antenna that could direct as much of our radio energy as possible towards this station, without radiating it in unwanted directions. This antenna would be *unidirectional*, favouring a single direction.

It turns out that in order to make a unidirectional antenna we need at least two radiating elements, which radiate in a specific phase relationship. This effect can be achieved in two ways. In *driven arrays*, the elements are all driven (that is, power is applied to them) in the correct phase relationship by a phasing network. In *parasitic arrays*, only one element is driven, but the antenna is designed so the driven element induces currents into the other *parasitic* elements.

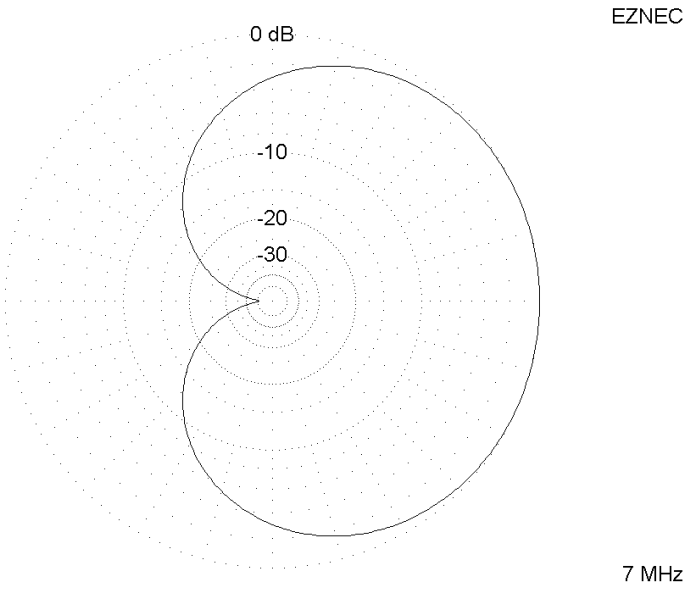
A simple driven array consists of two vertical elements spaced ¼ wavelength apart and driven 90º out of phase. This array is known as an *endfire array*, as shown in the view of this array as seen from above:



Endfire array, viewed from above

In this case element A (the left-hand element) is shown *leading* element B (the right-hand element) by 90º; conversely element B *lags* element A by 90º. Consider radio waves leaving B and heading in the direction of A. Since the elements are spaced ¼λ apart, it will take the radio waves a quarter of a cycle to get from B to A, by which time the phase of element A will have advanced by another quarter cycle, i.e. another 90º, in addition to the 90º phasing difference that already exists between the elements. So by the time radio waves from B get to A, they are 180º out of phase with the radio waves leaving A in the same direction (i.e. from right to left) and will cancel them out in that direction.

However, consider radio waves leaving A headed towards B. By the time they get to B, the phase of B will have advanced by 90º, making up for the 90º phase lag of the signal driving element B. So when radio waves from A reach B, they will be *in phase* with the radio waves radiating from element *B*, and so they will reinforce each other in the direction from left to right in the diagram. So waves heading from right to left will be cancelled; waves heading from left to right will be reinforced; and waves in different directions will be partially cancelled or partially reinforced according to some trigonometry that is too complex to go into here. The resulting pattern is shown below.



Azimuth radiation pattern of a two-element endfire array

This particular shape is called a "cardioid", meaning "heart-shaped", for obvious reasons. It is an example of a unidirectional radiation pattern. Note the excellent null of radiation from right to left, where the signals from the two antennas exactly cancel each other. By now, your cynical side should tell you that this null may not be all that deep, because of reflections against other objects in the vicinity.

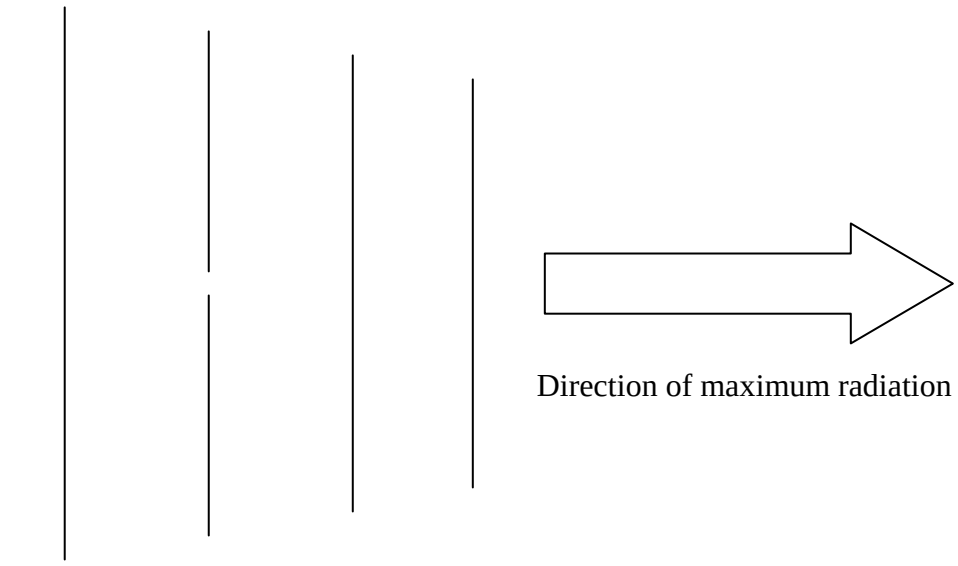
Although the driven array seems like a simple way of getting a unidirectional pattern, there are some complications in practice. In particular, it is not as simple as it looks to generate the necessary 90º phase difference between the two signals, since the antenna elements they are driving will have different impedances due to the interaction between them. However, they do have the advantage that by simply changing the phase relationship between the elements, the direction of the pattern can be reversed without having to physically rotate the antenna.

26.9 The Yagi

The full name for this antenna is the "Yagi-Uda Array", and it is named after Hidetsugu Yagi and his supervisor Shintaro Uda who invented it in 1926. History has been a bit unkind to Professor Uda, and the antenna is normally referred to just as the "Yagi". Mr Yagi is not entirely an innocent bystander in this history—he filed a patent that neglected to mention his mentor, and sold the patent to the Marconi company.

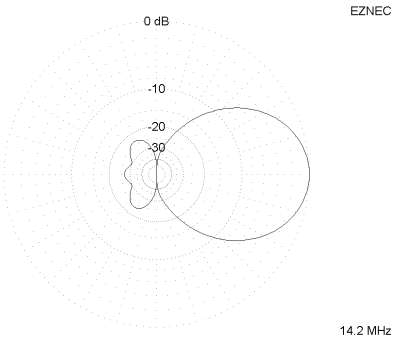
The Yagi consists of two or more half wavelength dipole elements, one of which—the *driven element*—is connected to the transmitter. The other element or elements are *parasitic*, meaning that currents are induced in them by induction from other elements. One of the parasitic elements will usually be a *reflector*, meaning an element that is on the other

side of the driven element from the desired direction of radiation, and the other elements (if any) will be directors, meaning elements placed in the desired direction of radiation. The reflector is usually slightly longer than the driven element in order to get the correct phasing, while directors are usually somewhat shorter than the driven element. The layout of a typical four-element Yagi is shown below:



A four-element Yagi array

The gap in the middle of the driven element is the feed-point, where the transmission line from the transmitter would be connected. The phasing between the elements is arranged by a careful choice of element lengths and separations so that radiation is reinforced in the desired direction and cancelled in other directions. The radiation pattern of a five-element Yagi is shown below.



Azimuth pattern of a five-element Yagi

Note how directional the pattern is – that is, how much radiation goes in the desired direction (from left to right) as opposed to other directions.

This directionality makes it an excellent antenna for amateur use, since as much radiation as possible can be beamed towards a distant receiver. Yagis are popular with amateurs on the higher HF bands, from perhaps 14 MHz upwards. They are usually mounted on towers with an electrically operated rotator that can point them in any desired direction. On lower bands, the large size can be a problem, but some astute amateurs operate Yagis on 7, 3,5 and even 1,8 MHz.

Other element configurations can also be used to make driven or parasitic arrays. For example, the cubical quad—usually called just the "quad" —is a parasitic array consisting of two or more quad loop elements. Quad and dipole elements can also be combined to form *Quagi* antennas. Directional antennas—including Yagis and quads—are often called "beam" antennas. Some complex Yagis include more than one driven element.

When coverage of vertical and horizontal polarisation is required, two sets of elements can be mounted on the same boom. One set provides vertical polarisation, while the other provides horizontal. The two antennas can also be fed together to provide circular polarisation, where the polarisation changes constantly in rotating fashion.

26.10 Reflector Antennas

A reflector antenna, such as a parabolic dish, is an attempt to increase the effective aperture of the antenna. A reflector is placed behind the feed point, which can be a simple dipole or something more elaborate such as a horn antenna mounted on a waveguide. The reflector then concentrates the transmitted energy from the feedpoint into a specific direction, or concentrates energy arriving from that direction into the feedpoint. From a transmission point of view, the functioning of a dish is not too dissimilar from the workings of an automotive headlamp reflector.

To produce appreciable gain, the dish must be many wavelengths in diameter. Dishes are therefore not practical below about 1 GHz. Below this frequency, Yagis or arrays are more efficient. However, a dish offers an attractive solution for UHF bands and above, as a single dish can house several feedpoints, covering several frequency bands with high gain. A 3 m dish is EME-capable above 1,2 GHz. At that frequency, the dish is over 12 wavelengths in diameter.

26.11 Antenna Gain

Antennas do not amplify (that is, increase the power of) signals. However, a directional antenna like a Yagi will radiate more of the available power in a particular desired direction. We speak about the *gain* of an antenna to mean the extent to which it is able to concentrate its radiation in a particular direction. Gain is expressed in dB, and is a measure of how much more power the antenna puts out in its most favoured direction, compared to the amount of energy that a reference antenna would put out in its most favoured direction.

Two different references are commonly used. The one is the half-wave dipole, in which case the unit of gain is dBd (decibels compared to a dipole). For example, if a Yagi was found to radiate four times as much power in its favoured direction as a dipole did (given equal transmitter power of course), it would have a gain of 6 dBd.

The other reference is the *isotropic radiator*. Isotropic means "radiating equally in all directions", not only horizontally (that would be an *omni-directional* antenna like a vertical) but all directions, including up and down. Although it would be very difficult to construct such an antenna, it is easy to calculate what the strength of its radiation would be if it was constructed, so by measuring the strength of the radiation from an actual antenna and comparing it with the calculated radiation strength from an isotropic antenna, we derive a gain figure in the units dBi (dB compared to an isotropic antenna).

In free space (that is, ignoring ground reflections) a dipole has a gain of 2,1 dBi. A gain figure for an antenna in free space in dBd can be simply converted to dBi by adding 2,1 dB. However, if ground reflections are taken into account, the gain of a dipole may be as much as 8,2 dBi. Be very careful in interpreting gain figures expressed in dBd, you need to know whether the reference dipole is at the same height as the antenna over the same ground medium, or if it is in free space.

Most antennas are reciprocal—they receive exactly like they receive. Directivity results in gain not only when transmitting, but also when receiving. An antenna with a gain of 3 dBd will convert signals from the desired direction into twice as much electrical power at the antenna terminals as a dipole would. This effect is not particularly useful by itself, since the limiting factor in receiver performance is usually atmospheric noise. However, directional antennas also do not respond to noise coming from directions other than the favoured directions, and this reduction in noise (both manmade interference and naturally occurring noise) gives directional antennas a significant advantage when receiving.

Be very wary of accepting published figures for antenna gain. It is very difficult to measure the actual gain of an antenna, and manufacturers know that purchase decisions are often based on the gain of an antenna, so they often use devious tricks to inflate the gain figures of their antennas. Commercial manufacturers are not solely to blame and designs published in amateur radio publications may also suffer from over-optimistic gain figures. Also be suspicious of anyone who quotes an antenna gain figure in "dB" without specifying the reference antenna. A number in decibels without a reference is meaningless, and anyone who does not understand this should not be considered an authoritative source for information about antenna gain!

26.12 Effective Isotropic Radiated Power

Assume that a dipole and a Yagi are both oriented so maximum radiation is in the direction of the same receiver, and that the Yagi has a gain of 6 dBd. Then if 100 W is applied to the dipole, and only 25 W to the Yagi, the signal strength in the desired direction will be identical because the gain of the Yagi will compensate for its lower input power.

Sometimes it is useful to express the amount of power radiated in the desired direction, irrespective of the actual power input or the gain of the antenna. This intensity can be expressed as the Effective Isotropic Radiated Power (EIRP), which is the power that you would need to supply to an isotropic antenna in order to radiate that much energy in the desired direction. EIRP can be calculated by multiplying the power actually applied to the antenna by the antenna gain with reference to an isotropic radiator.

For example, suppose our Yagi has a gain of 13 dBi, i.e. a gain of 20 with respect to an isotropic radiator, and the input power is 25 W. The EIRP is then 20 x 25 W = 500 W. The meaning is that to get the same amount of radiation in the favoured direction from an isotropic antenna, you would have to supply it with 500 W.

EIRP can be used to specify the power that you need to work a certain path. For example, the power required to operate a satellite might be specified as 1 kW EIRP. You could obtain this power by putting 400 W into an antenna with a gain of 4 dBi (such as a dipole), or 100 W into antenna with a gain of 10 dBi (such as a four-element Yagi), or 10 W into an antenna with a gain of 20 dBi (such as a long Yagi).

26.13 Efficiency

The efficiency of an antenna is the amount of power radiated (in all directions) as a percentage of the amount of power supplied to the antenna.

$$\eta = POut \div PIn$$

Efficiency is often expressed as a percentage. Just multiply the efficiency (which is a fraction between 0 and 1) by 100% to obtain the percentage efficiency.

Example: If 100 W is supplied to an antenna, but only 40 W is radiated as radio waves, the efficiency is 40%. The remaining energy is dissipated as heat in the antenna elements or the earth around the antenna.

Efficiency can be modelled simply as a resistor in series with the radiation resistance. A groundplane antenna with a feed resistance of 50 Ω has a radiation resistance of 36 Ω, which implies a loss resistance of 14 Ω. The radiated energy is 36/50 = 72% and the lost energy is 14/50 = 28%. The antenna is therefore 72% efficient.

Since the radiation resistance of an element drops rapidly as the length of the element is reduced below ¼λ (for a vertical) or ½λ (for a dipole), and the loss resistance remains relatively constant, antennas that are considerably shorter than these standard lengths may also be very inefficient. In practice, the situation may be even worse when inductive loading is added to compensate for the short elements, as the inductors would typically increase the loss resistance.

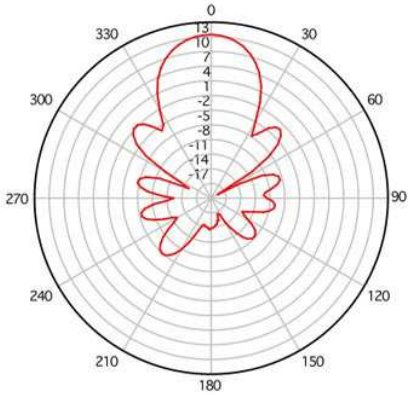
Some power being lost in head does not necessarily mean that the antenna won't work. Radio propagation variations are such that often very little radiated power is required to make a contact. When conditions are worse, the efficient antenna may just have the edge in making a contact.

26.14 Directivity as Opposed to Gain

A directional antenna does not necessarily have gain. If the antenna is lossy, like when it is made from high-resistance wire or when it is fed with an inefficient feeder, the antenna may be very directional, but with modest gain. A good example of this phenomenon is a Beverage receiving antenna. These antennas provide great directivity, providing much-improved reception of desired signals in the presence of atmospheric noise, but are very lossy. They are only useful for transmitting in very special circumstances. However, on the HF and VHF bands, most directional antennas provide equal benefit on transmit and receive because of their gain and directivity.

26.15 Other Performance Measures

The figure below shows the radiation pattern of a real commercial UHF antenna:



Azimuth pattern of a real antenna

Beam antenna performance is normally characterised by several numbers:

- **Gain**, normally in dBd or dBi. In this example, the gain seems to be 10 dB (reference not specified).
- **Beamwidth**, normally the angle between the -3 dB points on either side of the main axis. In the above figure, the beamwidth is perhaps 40° (i.e. 20° on either side of the main beam).

- **Front-to-back ratio** (F/B), about 27 dB in the above pattern. Forward radiation is at +10 dB, backwards at about -17 dB.
- **Sidelobe suppression** (SLS). A substantial lobe is radiated on either side of the main lobe, above 12 dB below the main lobe.

Depending on the application, some of these numbers could be more or less important. If relatively strong signals have to be received in the presence of interference, gain is not as important as a clean pattern (narrow beamwidth, good SLS and good FB). However, if ultra-weak signals have to be received on a clear frequency, gain is perhaps the most important parameter.

26.16 Stacking

Antennas can be stacked to provide narrower beams and more gain. Vertical stacking (one antenna above the other) retains horizontal beamwidth while reducing vertical beamwidth. Horizontal stacking (antennas side by side) reduces horizontal beamwidth while retaining vertical beamwidth. Stacking tends to degrade F/B and SLS.

All antennas in the array must be fed using feedlines. Some fancy footwork is required to match the impedance of each antenna to the feedline, and the many antenna feedlines to a single feedline for the transceiver.

On VHF and UHF, it is customary to stack many Yagis when high gain is required. Some EME stations use 32 or more Yagis, all pointing in the same direction, with the entire array being steerable in azimuth and often also in elevation.

26.17 Feedlines

In general, RF signals must be transported between the transceiver and the antennas in some way. Feedlines serve this purpose.

In feedlines, energy travels more slowly than in free space. The wavelength for a particular frequency is therefore also less in the feeder. The *velocity factor* for a specific feedline describes this phenomenon. Typical velocity factors for coaxial cables (see below) range from 0,5 to 0,8. For open-wire feeders, the velocity factor is close to 1. The velocity factor is very important when feedlines have to be cut to some fraction or multiple of a wavelength, such as when making phasing lines to connect several antennas together. The phasing line is much shorter than would be expected based on the free-space wavelength.

Feedlines basically come in two flavours: balanced and unbalanced.

Balanced feeders

Balanced feeders consist of two parallel wires. The characteristic impedance is determined by the ratio between wire diameter and wire spacing. They offer low loss and relatively low cost, but they are susceptible to interaction with the environment. They cannot be run close to window frames, fences or other metal structures. There are basically three types:

- **Open-wire feeders** are suspended in mid-air, normally by the occasional set of insulators. Old telephone lines can be regarded as open-wire feeders. Open wires are generally used over large distances in large antenna farms, such as with rhombic antennas. They have relatively low cost and very low loss. Its characteristic impedance can be tailored by adjusting the wire spacing, and is typically between 300 and 600 Ω.
- **Ladder line** consists of two parallel wires with occasional spacers of an insulating material. Typical spacing between wires is 100 mm, with spacers every 200 mm or so. Ladder line is convenient for feeding large dipoles and loops. Its characteristic impedance can be tailored by adjusting the wire spacing, and is typically between 300 and 600 Ω.

- **Twin-lead or ribbon** feedlines consist of two parallel wires, about 10 or 15 mm apart. The space between the two insulated wires is filled with a flat sheet of polyethelene. Twin-lead is generally used as a feeder for domestic broadcast receivers. It is rapidly disappearing, being replaced in most applications by coaxial cables. Its characteristic impedance is fixed, typically between 70 and 300 Ω.

Coaxial cables
A coaxial cable (or "coax"), where the core conductor runs inside a screen, is unbalanced. If everything works as advertised, the current flows in the core and the energy propagates in the insulating dielectric. There is no current on the outside of the cable. The cable should not radiate, and should not be susceptible to signals from outside. They can therefore run near metal structures and other cables.

However, if a balanced current is fed into the coaxial cable, current may flow on the braid and cause interference, both to and from the cable.

Coaxial cables come with an advertised characteristic impedance, determined by its geometry and its dielectric. For amateur radio purposes, this impedance is normally 50 Ω. Another common impedance, used in old Ethernet installations and in cable TV in other climes, is 72 Ω. In general, the impedance of the transmitter, the feedline and the antenna must all be the same. If they are not, matching is required.

Coaxial cables can become quite lossy, especially with long cables and at high frequencies.

Waveguides
Waveguides are unbalanced feeders that consist of enclosed tubes, most often rectangular in section. The energy runs inside the tube, in the form of current on the inside surface of the tube. They are usually used at microwave frequencies. They must be large in terms of wavelengths, making them impractical at HF and even VHF frequencies. Waveguides can handle very high power at relatively low loss.

As was mentioned in the section on dish antennas, a waveguide can be terminated with a flaring section, called a *horn antenna*, that works effectively as a feedpoint for a dish antenna.

26.18 Standing-Wave Ratio

The Voltage Standing-Wave Ratio (VSWR or simply SWR) describes how well the antenna is matched to the feedline. Ideally, the transmitter, the feedline and the antenna should all have the same characteristic impedance, to ensure maximum power transfer to the antenna.

If there is a difference between the antenna impedance and the line impedance, some of the energy will be reflected from the feedpoint, back into the feedline. The worse the difference is, the more energy is reflected. Depending on conditions at the transmitter end of the feedline and the loss of the feedline itself, some of the energy may be wasted. Most transmitters tend to cut back when the SWR is unfavourable, to protect themselves against damage by the reflected power. As a result, poor SWR could result in a weak transmitted signal.

The SWR is simply expressed as the ratio between the feedline characteristic impedance and the load impedance, expressed as a ratio greater than or equal to one. In the case of amateur antennas, the most common feedline is a 50 Ω coaxial cable. A perfect dipole will therefore have an SWR of 72/50 = 1,44:1, while a perfect ground plane antenna will have an SWR of 50/36 = 1,39:1. For practical purposes, an SWR of better than 1,5:1 is considered acceptable for most purposes, and 2:1 can generally be tolerated. Operating at

higher SWR should only be done with caution, as doing so could lead to feedline breakdown, transmitter damage and interference.

26.19 Baluns

An antenna that is symmetric w.r.t. the feedpoint such as a dipole or a full-wave loop is likely to be balanced, with exactly as much current flowing in the one half as in the other.

An asymmetric antenna such as a ground plane is probably not balanced.

Balanced feeders are, well, balanced. Coaxial cables are not. If we feed a balanced antenna with a coaxial cable, some current will tend to flow on the outside of the coaxial cable's screen, possibly causing interference or being susceptible to interference from outside.

The problem is solved with a *balun*, a balanced to unbalanced (get it?) transformer. A balun can be made from coaxial cable (such as the well-known bazooka), from a wire transformer wound on a toroid, or simply from lots of suitable toroids being slipped onto the outside shield of the cable.

26.20 Multiband Antennas

It is often useful to be able to use one antenna on several bands. With eight amateur band allocations in the HF region alone (3 to 30 MHz), very few amateurs are able to put up a separate antenna for every band. There are two major considerations with multiband antennas: radiation pattern and impedance matching.

This section will not consider the radiation pattern of multiband antennas in any detail, except to note that the patterns may be very different on the different bands and that this is an important consideration when designing multiband directional antennas.

In the section on SWR, we showed that dipoles and verticals are easily matched to 50 Ω cables. However, on other bands the impedance of these antennas may be considerably different from what it is on the design frequency. For example, if a dipole cut for 14 MHz is used on 28 MHz, the impedance will be more like 5 kΩ, with a resulting SWR of 10:1!

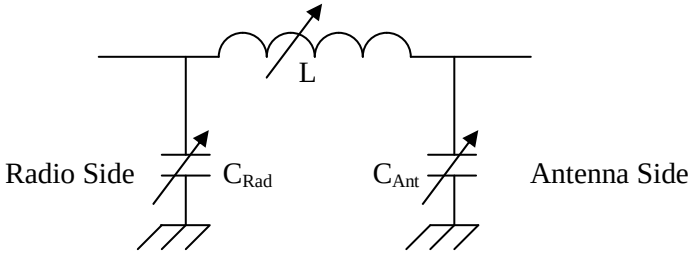
There are four common solutions to the impedance matching problem:

Antenna Tuning Unit

An Antenna Tuning Unit (ATU)—also known as an "antenna coupler" or "Transmatch"—to match just about any impedance to 50 Ω. Use a simple single-element antenna like a dipole, and just match it on any frequency of interest using the ATU. The ATU normally contains an inductor and two capacitors that can be adjusted until a good match is achieved. Classic ATUs feature three rotating knobs that have to be twiddled interactively, but automatic ATUs are now available. These automatic models will match the antenna quickly, based on a sample of the RF transmitted into them.

ATUs come in two basic flavours: Pi and T networks. The letters refer to the basic layout of components in the ATU.

A Pi network is named for its shape, resembling a Greek capital P (Π):



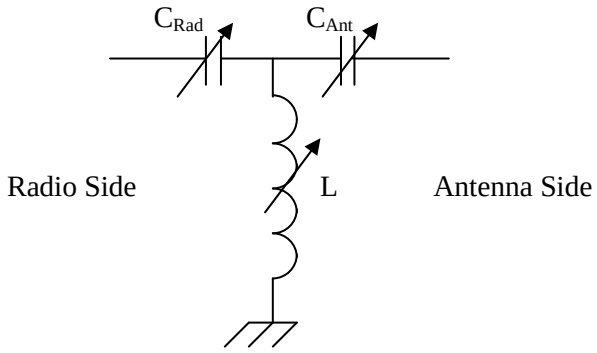
Pi network ATU

The user (or an automatic controller) would adjust the two capacitors and the inductor until the best match is found. The network is unbalanced. Similar networks can be built with another inductor in the lower rail, for balanced antennas.

This configuration of the Pi network is clearly a low-pass filter. In this configuration, it is a typical output network for transmitters. Tube-type linear amplifiers typically have adjustable Pi networks to match the tube to the antenna, requiring adjustment by the operator whenever the frequency changes. More modern amplifiers, including most solid-state units, do not require user adjustment. They either use a number of preselected Pi networks, or they automatically adjust the networks for the best match during use.

A Pi network could also be made with a capacitor in the top rail and inductors on the input and output. In this configuration, the Pi network would be a high-pass filter. This configuration is less useful in transmitter outputs, as it is more expensive due to the two inductors and does not perform a low-pass function.

A T network is named for its shape, that resembles a capital T:



T network ATU

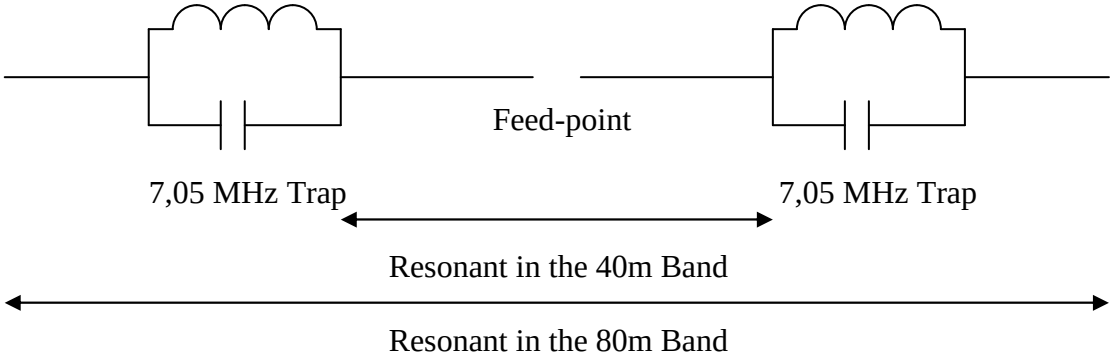
For this solution to work efficiently, either the ATU must be mounted close to the antenna, or a low-loss high-impedance feeder (open wire line) must be used to connect the antenna to the ATU. You cannot expect any degree of efficiency if you run 50 m of 50 Ω coax to the antenna with the ATU on the transceiver side (i.e. not on the antenna side of the coax connection).

Fan Dipole

One can wire several elements in parallel, provided that for each frequency band, one and only one element has a low, purely resistive, impedance. For example, three dipoles cut for the 20, 15 and 10-metre bands (14, 21 and 28 MHz respectively) could be connected in parallel (i.e. all share the same feed-point). The elements are normally spaced outwards by a few degrees, forming a fan-like structure. This structure is known as a *fan dipole*. If a fan dipole is fed with a 14 MHz signal, only the 20 m dipole will have a low impedance, so

almost all the energy will go into that dipole. Similarly, if it is fed at 21 or 28 MHz, only the dipole for that frequency band will have low impedance, and almost all the energy will go to that one.

Traps
Traps can be used to effectively shorten the antenna at higher frequencies. A trap is simply a parallel tuned circuit that appears as a high impedance at its resonant frequency, and as a low impedance at other frequencies. The diagram below shows a trap antenna for use on the 80 m and 40 m bands.



A Trap Dipole for 40m and 80m

The inner section of the antenna would be resonant in the 40 m band (7,0 to 7,2 MHz). It is essentially a simple dipole for 7 MHz. The traps are also resonant in this frequency range, presenting a high impedance and effectively disconnecting the outer sections of the antenna at these frequencies. However, when the antenna is fed at a frequency in the 80 m band (around 3,5 MHz), the traps are low impedance, acting like inductors. They therefore allow the whole antenna to be active, including the inductance of the traps, and the total length is designed to be resonant in the 80 m band. The total length of the antenna is less than would be required for a trapless (single band) 80 m antenna.

Trap antennas can be designed for more than two bands by adding another trap (or pair of traps in the case of a dipole) for each additional band. However, since each trap will have some loss, a trap antenna for many bands might be quite inefficient on the lower bands, where antenna current is flowing through multiple traps.

Multiple Resonances
Finally, you can make use of the fact that most antennas are naturally resonant on more than one frequency. For example, open-ended antennas like dipoles and verticals are resonant on odd harmonics of the fundamental frequency—that is, 3, 5 and 7 times the original design frequency. For example, a dipole designed for the 40 m (7 MHz) band may also be resonant on the 15 m (21 MHz) band, which is the third harmonic of the design frequency. Full-wavelength closed loop antennas like the quad antenna are resonant on all harmonics, so in theory an 80 m quad loop should also be resonant in the 40, 20 and 10 m bands. Matching the loop may not be so easy though, as the resonant frequencies may not be precisely where you need them, and that it is difficult to tune them to the desired frequency without changing the other resonant frequencies! Also, don't forget that the radiation pattern may not be what you were hoping for on some of the other bands.

26.21 The Log-Periodic Array

A beam antenna that covers a wide frequency range is the log-periodic array (LPA), alternatively called a log-periodic dipole array (LPDA). The array consists of many dipoles, each resonant on a different frequency. The dipoles are all fed together via a common

feedline. On any particular frequency, one element is approximately resonant and acts as driver, while longer elements act as reflectors and shorter elements act as directors. Depending on the ratio of the length of the shortest dipole to that of the longest, the array can cover a frequency range of 10:1 or more.

The figure shows a fixed-direction LPA made from wire. It seems to cover approximately a 3:1 frequency ratio with relatively high gain.

Log-periodic array

Relatively few amateurs use LPAs, as single-band or multi-resonant antennas can easily provide better performance on the few bands we occupy. Commercial or military users that use a wide variety of frequencies may prefer the flexibility afforded by the LPA.

26.22 Making Practical Antennas

As you will have read earlier in this text, antennas provide a great avenue for experimentation and home construction. You can save considerable money and gain considerable satisfaction from making your own antennas.

For the HF bands, antenna construction is easy. Accuracy is not very critical, as wavelengths are in the order of many metres, and a few mm here and there will hardly be noticed.

Many amateurs use existing supports such as trees and buildings to string their antennas. A catapult or bow and arrow can be used to sling a monofilament fishing line across a branch at a decent height. The line can then be used to raise a stronger cord that can be used to hoist the antenna.

A dipole can be made simply by calculating a quarter wavelength in free space, reducing the dimensions by about 5% to compensate for velocity factor, then cutting the two legs of the dipole from thin uninsulated electrician's earthing wire. Leave about 10% spare on either end, just in case it doesn't resonate where you expect. You can wrap the spare length of the wire back onto itself. Cheap plastic egg insulators made for electric fences are available from your hardware dealer. Use these to insulate the two halves of the dipole from one another and the ends of the dipole from the support lines. Now wind a choke from about 10 turns of coax cable near the feedpoint to reduce current on the braid. Take care not to wind the choke too tightly, or you may damage the cable. A coil diameter of 10 times the coax diameter is about right. The choke can be secured with cable ties or binding wire. Now hoist the antenna into position and try it out. If you have to, you can shorten or lengthen the dipole slightly, using the spare wire you so judiciously left when you cut the wire to length.

Don't get hung up on fancy simulations. Your antenna may not be perfect, but it is likely to provide adequate results. Remember: The worst antenna above your house still works better than the best simulation in your computer!

If you want to progress beyond a basic antenna, you can consult any of the dozens of good antenna books and amateur radio magazines for great construction articles.

Summary

Antennas convert electrical energy into radio waves that can be radiated long distances. Electromagnetic waves consist of a magnetic field and an electric field that are both at right angles to each other, and at right angles to the direction of propagation of the wave. The *polarisation* of radio waves depends on the orientation of the electric field—if the electric field is horizontal, the wave is said to be *horizontally polarised* and if it is vertical, the wave is *vertically polarised*. Antennas do not respond well to radio signals with the "wrong" polarisation.

Every antenna has a *capture area*, which represents the cross-section from which the antenna can extract energy from an incoming radio wave.

The half-wave dipole consists of a centre-fed ½λ element. Its radiation resistance is approximately 72 Ω. In free space, a dipole has a doughnut-shaped radiation pattern. A horizontal dipole over ground has a bi-directional pattern similar to a figure "8". Dipoles can be deformed to fit into available space, or erected from a single support as an Inverted V.

The ¼ λ vertical consists of a vertical ¼ λ radiator that is fed either against ground or against a "ground-plane". It has an omni-directional pattern, radiating equally in all azimuth directions.

A unidirectional pattern can be achieved using a multi-element antenna, either a *phased array* where each element is individually fed from a phasing network, or a *parasitic array*, where the transmitter feeds only one element and the others are excited by inductive coupling from other elements. The most common parasitic array is the Yagi, which consists of two or more elements—the driven element, a reflector, and one or more directors. All elements are approximately ½ λ long.

A dish or reflector antenna uses a dish-shaped reflector to concentrate the energy from the feed point in a specific direction. Dishes are mostly useful on UHF and above, as they must be many wavelengths in diameter to provide useful gain. The feedpoint can consist of a relatively simple antenna such as a dipole or horn.

The gain of antenna expresses how much power is radiated in the most favoured direction, compared with some reference antenna. Gain can be specified in dBd (dB compared to a dipole) or in dBi (dB compared to an isotropic radiator). The efficiency of antenna is the amount of power radiated as a percentage of the total power applied to the antenna. The Effective Isotropic Radiated Power (EIRP) is the power fed to the antenna multiplied by the gain of the antenna with respect to an isotropic radiator.

Identical antennas can be stacked and fed together to provide improved performance.

Other important parameters of antennas (apart from gain) include beamwidth, sidelobe suppression and front-to-back ratio (F/B).

Feedlines come in two varieties: Balanced and unbalanced. Balanced feeders offer low loss and simplicity under some conditions, but are prone to interaction with nearby structures. Their impedance can be shaped by changing the wire spacing. Coaxial cables have a fixed characteristic impedance, and are more or less impervious to outside influences if fed properly. They can become very lossy over long distances or at high frequencies.

Energy travels more slowly in feedlines than in free space. As a result, the wavelength is less in a feedline than one would expect. The ratio between the actual wavelength and the free-space wavelength is known as the *velocity factor*. For coaxial cables, this factor is around 0,7.

Standing-wave ratio (SWR) indicates how well an antenna is matched to the feedline, and is expressed as a ratio of 1:1 or greater. An SWR of 1,5:1 at the transmitter is considered good enough, and up to 2:1 can be tolerated.

A balun (balanced to unbalanced transformer) should be used to match balanced loads to unbalanced coaxial cables to prevent interference from currents flowing in the braid.

An antenna may be impedance-matched on multiple bands by using an antenna tuner, by feeding multiple elements in parallel, by using traps or by taking advantage of naturally occurring harmonic resonances.

The Log-Periodic Array (LPA) provides wide frequency coverage with modest gain.

HF antennas provide an excellent starting point for antenna experimentation. All you need are a few egg insulators, some coax cable and some bare copper wire.

Revision Questions

1 Electromagnetic waves are created by:

a. The alternating RF currents in an antenna.
b. Magnetic solenoids.
c. Audio loudspeakers.
d. DC voltages.

2 In electromagnetic radiation, which of the following is true?

a. E and H are at 180° to each other.
b. E, H and the direction of propagation are all at right angles to each other.
c. The angle between E and H is 0°.
d. The velocity of propagation is at 180° to the E field but in line with the H field.

3 In order to radiate, an electromagnetic wave must have:

a. E Field only.
b. H Field only.
c. E and H Field
d. Air to travel in.

4 Polarisation of an electromagnetic wave is fixed by:

a. The direction of the H field.
b. The direction of propagation.
c. By an anti-phase signal.
d. The orientation of the transmitting antenna.

5 The wavelength of a signal of 100 MHz in free space is:

a. 30 mm
b. 0,3 m
c. 3,0 m
d. 30 m

6 When an antenna is well-matched to the feedline at the transmitted frequency:

a. Maximum power will be reflected.
b. A good SWR will be obtained.
c. The SWR wil be poor.
d. An SWR reading will be meaningless.

7 What do the terms vertical and horizontal, as applied to wave polarisation, refer to?

a. Orientation of the electric field.
b. Orientation of the magnetic field.
c. Orientation of the charge particles in the propagation medium.
d. Launching angle of the wave with respect to the earth's surface.

8 What radiation pattern does an ideal half-wave dipole have if it is installed parallel to the earth?

a. It radiates well in both horizontal directions, perpendicular to the dipole.
b. It radiates poorly in a horizontal direction and parallel to the dipole.
c. It radiates equally well in all horizontal directions.
d. It radiates poorly in all horizontal directions, but it radiates well in a vertical direction.

9 How does proximity to the ground affect the radiation pattern of a horizontal dipole antenna?

a. If the antenna is too far from the ground, the pattern becomes unpredictable.
b. If the antenna is less than half a wavelength from the ground, reflected radio waves from the ground distort the antenna’s radiation pattern.
c. A dipole antenna’s radiation pattern is unaffected by its distance from the ground.
d. If the antenna is less than half a wavelength from the ground, radiation off the ends of the wire is reduced.

10 Which kind of antenna would best enhance a signal from a particular direction, while rejecting interfering signals from other directions?

a. A monopole antenna.
b. An isotropic antenna.
c. A vertical antenna.
d. A beam antenna.

11 What is a directional antenna?

a. An antenna whose parasitic elements are all constructed to be directors.
b. An antenna that radiates in direct line-of-sight propagation, but not skywave or skip propagation.
c. An antenna permanently mounted so as to radiate in only one direction.
d. An antenna that radiates more strongly in some directions than others.

12 What is the purpose of an antenna matching circuit?

a. To measure the impedance of the antenna.
b. To compare the radiation patterns of two antennas.
c. To measure the SWR of an antenna.
d. To match impedances within the antenna systems.

13 When will a power source deliver maximum output?

a. When the impedance of the load is equal to the impedance of the source.
b. When the SWR has reached a maximum value.
c. When the power supply fuse rating equals the primary winding current.
d. When air wound transformers are used instead of iron core transformers.

14 What is the wavelength of a 100 MHz signal in RG213 coaxial cable?

a. About 2 m
b. About 3 m
c. About 4 m
d. About 6 m

15 What is a Yagi antenna?

a. Half-wavelength elements stacked vertically and exited in phase.
b. Quarter-wavelength elements arranged horizontally and excited out of phase.
c. A half-wavelength linear driven element (or elements) with parasitically excited parallel linear elements.
d. Quarter-wavelength, triangular loop elements.

16 Why is a Yagi antenna often used for amateur radio communications on the 20 m amateur band?

a. It provides excellent omniderectional coverage in the horizontal plane.
b. It is smaller, less expensive and easier to erect than a dipole or vertical antenna.
c. It discriminates against interference from other stations off to the side or behind.
d. It provides the highest possible angle of radiation for the HF bands.

17 Choose a physical description of the radiating elements of a horizontally polarised Yagi antenna.

a. Two or more straight, parallel elements arranged in the horizontal plane.
b. Vertically stacked square or circular loops arranged in parallel horizontal planes.
c. Two or more wire loops arranged in parallel vertical planes.
d. A vertical radiator arranged in the centre of an effective RF ground plane.

18 What is the name of the parasitic beam antenna using two or more straight metal elements arranged physically parallel to each other?

a. A quad antenna.
b. A delta loop antenna.
c. A zepp antenna.
d. A Yagi antenna.

19 How many driven elements does a simple Yagi antenna have?

a. None; they are all parasitic.
b. One.
c. Two.
d. All elements are driven.

20 What kind of antenna array is composed of a square or diamond-shaped full-wave closed loop driven element with one or more parasitic loops parallel to the first one?

a. Dual rhombic.
b. Cubical quad.
c. Stacked yagi.
d. Delta loop.

21 An amateur finds that a Yagi just does not provide enough signal to access a distant repeater. A possible solution is to:

a. Use an LPA.
b. Use a dipole.
c. Use a Ground Plane antenna.
d. Use two identical Yagis on the same tower.

22 What is the polarisation of the signal from a half-wavelength antenna which has elements perpendicular to the earth's surface:

a. Circularly polarised waves.
b. Horizontally polarised waves.
c. Parabolically polarised waves.
d. Vertically polarised waves.

23 A dipole with two sets of traps will allow operation on:

a. One band.
b. Two bands.
c. Three bands.
d. All bands.

24 A folded dipole has an approximate impedance of:

a. 50 Ω
b. 72 Ω
c. 150 Ω
d. 300 Ω

25 A vertical antenna relies upon:

a. A good earth and ground connection.
b. No earthing.
c. A sensitive receiver.
d. The D layer.

26 The term Zepp, Yagi, Quad and Log Periodic refer to:

a. Oscillators.
b. Transistors.
c. Antennas.
d. Diodes.

Chapter 27: Propagation

Propagation is the process by which radio waves get from the antenna of the transmitter to the antenna of a distant receiver. This chapter introduces the different propagation modes used by amateurs.

27.1 Frequency Bands

For this discussion, you must remember the terminology from an earlier chapter:

Range		Wavelength		Frequency		Amateur bands
		From	To	From	To	
HF	High frequency	3 MHz	30 MHz	100 m	10 m	80 to 10 m (9 bands)
VHF	Very high frequency	30 MHz	300 MHz	10 m	1 m	6 m, 2 m
UHF	Ultra-high frequency	300 MHz	3 GHz	1 m	100 mm	70 to 23 cm

Let's add some more rather fuzzy terminology: When we refer to *high bands*, we're talking about the high HF and low VHF spectrum, perhaps the 15, 12, 10 and 6 m bands. The *low bands* are in the low HF and upper MF spectrum: 160, 80, 60 and perhaps even 40 m. The *mid-bands* are those between these two groups, perhaps starting with 40 m and going up to 17 m.

Another term that we will use regularly is DX. DX is an old telegraphy abbreviation for "distance", and is used to describe long-distance contacts. The definition is a little flexible. On the high bands, long-distance contacts are easy and a contact to central Africa might not be regarded as DX. On the low bands, with less effective antennas and higher noise, or on VHF, with mostly line-of-sight communications, the same contact might be construed as great DX.

27.2 Direct Wave (Line of Sight) Propagation

Electromagnetic radiation generally travels in straight lines, so if radio waves can travel straight from the transmitting antenna to the receiving antenna without being blocked by obstructions, communication is possible. This simplest form of propagation is known as "direct wave" propagation. It is also called "line-of-sight" propagation although this term is a bit misleading since some things that block light, such as a wooden structure, will allow radio waves to pass through.

Direct wave propagation affects all frequencies. The possible range depends on the terrain and the height of the antennas. Over flat terrain, with both antennas 10 m high, the range of direct wave propagation is about 20 km. However, hilly terrain can be used to good effect by placing one of the antennas on top of a hill where it can be "seen" from much further away. VHF repeaters are usually located on high sites, since they rely on direct wave propagation. They can achieve wide coverage at distances of perhaps 100 km.

Both horizontally polarised and vertically polarised waves propagate equally well over line of sight. However, because this form of propagation retains the original polarisation of the wave, it is important to ensure that both transmitting and receiving antennas have the same polarisation.

27.3 Ground Wave Propagation

Low and medium frequencies *refract* around the surface of the earth. Refraction is caused by the nearby ground slowing the radio wave down slightly, causing it to bend towards the ground. Because the ground itself is bending with the curvature of the earth, the effect is that the ground wave follows the earth's surface. Refraction is most pronounced at lower

frequencies, so this effect is most significant in the LF and MF bands. It is present but less effective in the HF bands and absent at VHF and above.

MF (or "medium wave") AM commercial broadcast stations can be received up to 100 km or so away from the transmitter. The medium wave broadcast band (530 kHz to 1,6 MHz) are low enough for good ground wave propagation to occur. However, commercial FM transmitters use VHF frequencies (88 to 108 MHz), which are only propagated by direct wave, so they are only usable within 10 to 50 km of the transmitter, depending on terrain.

The same ground interactions that allow the wave to refract around the curvature of the earth also attenuate it, limiting the range of ground wave propagation to a few hundred kilometres, depending on the power of the transmitter.

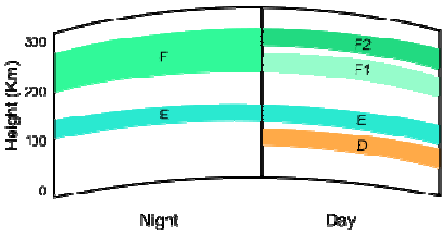
27.4 The Atmosphere

The atmosphere consists of three layers: troposphere, stratosphere and ionosphere. The troposphere extends from the surface of the earth to a height of about 10 km. It is the area where most of the weather we are familiar with happens. The stratosphere extends from 10 km above the surface of the earth to approximately 50 km. In this region, the temperature and humidity remain relatively constant. The stratosphere has little effect on radio wave propagation.

The ionosphere is that part of the upper atmosphere where free electrons occur in sufficient density to have an appreciable influence on the propagation of radio frequency electromagnetic waves. It extends from approximately 50 km to 800 km above the earth's surface. In the ionosphere, high-energy solar radiation (x-rays, ultraviolet radiation and particles from the "solar wind") strips electrons from some gas molecules, leaving positively charged ions and free electrons. This pea-soup of particles is known as a *plasma*.

The ionosphere is divided into four layers. The D layer, which extends from 50 to 90 km above the surface, is only present during daylight hours. As soon as the sun's ionising radiation is no longer present, electrons and ions rapidly recombine to form neutral (un-ionised) gas, and the D layer disappears. The principal effect of the D layer is to absorb radio waves. Although some absorption takes place at all frequencies, the amount of absorption decreases with the square of the frequency, so it affects low frequencies much more than high frequencies.

The upper three layers have a different effect. Instead of simply absorbing radio waves, they bend the waves by refraction. If a wave is bent sufficiently, it may return to earth a considerable distance from the transmitter, almost as though it had been reflected off the ionosphere. The amount of refraction (bending) depends on frequency, and is more pronounced at lower frequencies. The upper layers are the E layer, which extends from about 90 to 150 km above the surface; the F_1 layer, which extends from about 150 to 180 km; and the F_2 layer, which extends from about 180 to 300 km or higher. At night, the E layer dissipates while the F_1 and F_2 layers combine to form a single F layer that is less strongly ionised than during the daytime.



Ionospheric layers

You may also encounter other names for these layers, with the E layer being known as the Kennelly-Heaviside layer and the F layer being known as the Appleton-Barnett layer.

27.5 Sky Wave (Ionospheric) Propagation

During the daytime, the D layer will absorb low frequencies, but higher frequencies will penetrate the D layer (albeit with some attenuation) and can be refracted back to earth by the E, F_1 or F_2 layers.

Even higher frequencies will not be refracted sufficiently by the E, F_1 and F_2 layers and will continue straight out into space instead of being returned to earth. Refraction by the F_2 layer (or at night by the single F layer) is responsible for most long-distance HF communication.

At night, the D layer dissipates almost immediately and the E layer more gradually, while the F_1 and F_2 layers combine to form a single less strongly ionised layer. Now that there is no D layer to absorb low frequencies, they can be reflected from the F layer and travel long distances. However, high frequencies are not be refracted sufficiently by the more weakly ionised nighttime F layer, so they will be lost into space.

Note that fairly high frequencies may still be usable well after local sunset. Because a plasma is a poor conductor, it takes a considerable time for ionisation levels in the F_1 and F_2 layers to decrease to their nighttime values. Also, because these layers are high above the surface of the earth, they will be illuminated for some time after local sunset. And finally, for paths from east to west, the point where the waves need to be refracted will be located some distance to the west of the transmitter, where the sun may still be shining.

This process of being refracted from the ionosphere is also known as “skip”. The maximum skip distance for the E layer is around 2500 km, and about 5000 km for the F layer. Longer paths may be achieved by *multi-hop propagation*, where the refracted signal bounces off the surface of the earth back to the ionosphere and is refracted back to earth again. Up to ten such hops can happen under exceptional circumstances, providing propagation to points anywhere on earth, maybe even via the long path.

There is normally a range of distances in which a signal cannot be heard, as the receiver is too close to the transmitter to receive skip and too far for direct or ground wave. This area is known as the *skip zone*. In the skip zone, no or weak signals can be heard. On the high bands, the skip zone typically extends from about 100 to 1000 km.

The highest frequency that can be used on a particular path (i.e. for communication between two particular places at a particular time) is called the Maximum Usable Frequency (MUF) for that path. It is dependent mainly on the degree of ionisation present in the ionosphere. Increasing the EIRP won’t help, because if the frequency is above the path MUF, the additional power will just be radiated into space.

A signal may arrive at its destination via two or more paths that happen to be open at the same time. The ionosphere is constantly changing, so *multipath propagation* is also constantly changing. Signals arrive with different time delays, resulting in phase differences. In exceptional cases, when signals arrive via short and long path at the same time, the difference could be as much as 150 ms. The signals may add or subtract, depending on the phase difference, resulting in enhancement or destruction of the signal. The result is a constant change in amplitude known as *fading*. Fading could vary considerably from a short period (fractions of a second) to long periods (many minutes) and could be tens of dB deep.

The lowest frequency that can be used for a path is called the Lowest Usable Frequency (LUF). The LUF depends not only on the amount of ionisation, but also on the amount of atmospheric and man-made noise present at the receiver and on the EIRP of the signal. The main consideration is whether the transmitted signal, after being attenuated by the D layer, is still sufficiently strong to be heard above the noise, so additional power will help to decrease the LUF.

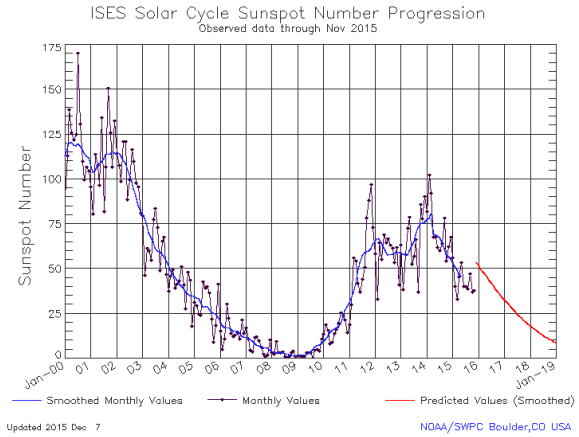
D-layer absorption and atmospheric noise both decrease as the frequency increases, so the best propagation is usually to be found just below the MUF. The best frequency, producing best communications for the path, is called the Optimal Traffic Frequency (OTF) or by its French name, FOT.

The critical frequency is the highest frequency that can be radiated vertically upwards and still return to earth. The critical frequency is only of indirect interest to amateurs, since usually one is not hoping to bounce a signal vertically off the ionosphere and have it return to the house next door. The critical frequency can be measured relatively easily by scientists, and is often the best indicator of ionospheric conditions in the immediate vicinity. When the critical frequency is high, MUFs for paths in that vicinity are likely to be high as well, and when it is low, path MUFs will be low.

The actual MUF for a particular path may be well above the critical frequency, since waves radiated at a shallow angle may be returned to earth when waves radiated vertically upwards are not. Think of how you can skip a flat pebble on a pond if the pebble arrives at a flat enough angle. Radio waves behave much the same way—if they hit the ionosphere almost vertically, they will probably penetrate, but if they arrive at a shallow angle, they may well return to earth.

MUFs and LUFs depend on the extent of ionisation in the ionosphere. This ionisation varies with the time of day; with the season and with the amount of solar activity. Solar activity follows cycles approximately 11 years long. Every 11 years there is a solar maximum, when high levels of solar activity generate intense ionisation and MUFs can extend well above 50 MHz during daylight hours, giving great long-distance openings on the 6 m and 10 m amateur bands. About seven or eight years later there will be a solar minimum, where MUFs may be under 20 MHz. The 6 m and 10 m bands will be mostly dead, while propagation on the low bands (160 m and 80 m) will be better than usual.

Solar Cycle 24 peaked in 2015. Solar cycles have been numbered since 1755, when regular observations of sunspots started being recorded. The diagram shows the sunspot number by day. Cycle 23 peaked around 2001 at an SN of about 120. Cycle 24 peaked twice. The highest peak was in 2015 at an SN of about 80, leaving a lot of amateurs very disappointed. Cycle 24 is said to have been the worst in the history of radio:



Solar cycle 24

Solar activity is measured in two different units: the *sunspot number* (SN) and the *solar flux index* (SFI). The two correlate very well, allowing modern flux measurements to be compared to previous cycles when only optical observations of sunspots were available. Other important parameters include A and K, both of which give an indication of how stormy the solar surface is.

The sunspot cycle affects the optimum frequencies for communication. At the sunspot maximum, daylight DX (long distance) communication will typically use the high bands, with the other bands being used only at night. At a sunspot minimum, daytime DX bands will typically be the mid-bands, with the low bands being used for DX at night.

Of course propagation does not depend only on conditions at the transmitter, but equally on conditions at the receiver and along the whole path. For example, it would not make much sense to try to contact stations in the USA on a daylight band at 10:00 local time in South Africa, since at that time it is only 03:00 local time on the east coast of the USA. However, at 10:00 local time in South Africa, the daylight bands might work well for contacting Japanese stations, since then it is 17:00 in Tokyo. In general, paths that are either all daytime or all-night are the easiest. Mixed daylight and nighttime paths can be difficult since frequencies that work on one side of the link won't work on the other side, and *vice versa*. However, remember that "daytime" conditions may persist for several hours after local sunset, and even later for east-west paths.

In general, radio waves travel along the great circle path between their source and destination. This path that represents the shortest distance between two points on the surface of the earth, without actually going *through* the earth! However, there are *two* great circle paths between any points on earth—a short path and a long path. For example, the short path to the USA from South Africa is to the northwest; while the long path is in the opposite direction, to the southeast, and travels around the world, crossing Australia and the Pacific before finally ending up at the west coast of America. Radio propagation can be via either long or short paths, depending on the conditions along each path. In the late afternoon, it may be possible to communicate with a station in the USA via short path on the high bands and via long path on the low bands, at the same time.

Ionospheric propagation achieves the longest distances when the takeoff angle of the radio signal (the vertical angle of the signal above the horizon) is small. We have already mentioned the fact that signals that arrive at a shallow angle to the ionosphere are much more likely to be reflected. In addition, flatter hops cover more distance, requiring less hops to get to the destination. Generally, a low takeoff angle is required for DX. In the case of horizontally polarised antennas, low takeoff angles are achieved at great heights, generally at least one wavelength above ground.

At dawn and dusk, as the transition from day to night or night to day takes place, the transition zone between daytime and nighttime conditions causes tilting of the ionospheric layers that may provide fleeting propagation over huge distances. Low band DXers spend most sunrise and sunset periods listening and calling to exploit these *grey-line* openings.

Ionospheric propagation is most common on the MF (at night only) and HF bands. It occurs occasionally in the low VHF region, for example the 6 m amateur band, which is also called the "magic band" because "skip", albeit infrequent, can travel great distances with very little power.

27.6 Exotic Ionospheric Propagation Modes

Sporadic E Propagation

A form of ionospheric propagation that affects VHF transmissions is known as Sporadic-E (E_S) This propagation mode consists of the refraction of VHF frequencies from small patches of intense ionisation in the E layer. The cause of these intensely ionised areas is not well understood and they appear unpredictably (hence "sporadic"). E_S is most frequently during summer daylight hours and may last for several hours. E_S is usable on frequencies from 28 to 220 MHz and signal strengths are often very strong, with low-power transmitters being heard hundreds or thousands of kilometres away. Path distances may exceed the 2500 km maximum for single-hop E layer skip, indicating that either multiple hops or some form of ionospheric ducting is present.

Backscatter

When ionospheric propagation takes place, some of the signal after the first hop is scattered back in its direction of origin. The signal then reflects from the ionosphere, landing up in the general area where it originated. The signal is normally relatively weak and would go completely unnoticed by many operators, but can be copied with some effort. *Backscatter* sometimes allows contacts in the skip zone, and is the most likely mode for countrywide high-band contacts.

Meteor Scatter

Meteors entering the earth's atmosphere leave a trail of ionised gas in their path that can refract VHF signals. The ionisation typically only lasts for a few seconds or tens of seconds before the electrons and ions recombine. Very fast speech or Morse code, or specialised digital modes like JT6M and FSK441, is needed to take advantage of meteor scatter. Typical meteor scatter bands are 6 m and 2 m.

Auroral Scatter

When the sun is unusually active and solar debris has accumulated at the magnetic poles, aurora forms. Aurora absorbs HF signals, but intense aurora can reflect VHF and UHF signals. Because the reflective medium is heterogeneous and constantly changing, the reflected signals experience Doppler shift and phase dispersion. CW signals can sound rough and raspy. Auroral scatter is only accessible to stations close to the polar regions, mostly in northern Europe, Asia and North America. Distances of thousands of km can be achieved.

27.7 Tropospheric Bending, Scatter and Ducting

Some bending (refraction) of VHF and UHF radio waves occurs in the troposphere, which increases the "radio horizon" (the distance over which radio waves can propagate without reflection or scattering) by about 15% compared to the visual horizon. A simple way in which this effect is modelled is to assume that the earth is actually about one-third larger than it really is when working out if coverage is possible. This approach is known as the "*four-thirds earth*" model.

Temperature and humidity irregularities within the troposphere (the lower 10 km of the atmosphere) can reflect VHF and UHF signals over a distance of from 100 to 500 km or so. The reflections are usually fairly weak, so reasonable EIRP (either a high power transmitter or an antenna with gain) are required. However, unlike meteor scatter, tropo-scatter is long lived, so it is possible to use standard modes like CW and SSB for tropo-scatter work. FM is not recommended for weak-signal work, as it requires more power for an intelligible signal than either CW or SSB.

In tropospheric ducting, VHF signals are "trapped" between an inversion layer and the ground or between two inversion layers and may travel thousands of kilometres with little attenuation. These openings are found mostly along the coastline and are difficult to predict.

27.8 Earth Moon Earth (EME)

EME involves bouncing signals off the moon to some distant location on earth. On VHF bands and above, EME is the primary mode for intercontinental DX contacts, as ionospheric propagation is exceedingly rare.

The losses are tremendous, with a path loss of 250 dB being typical. Only about 0,000 000 000 000 000 000 001% of the transmitted signal returns to earth! An EME station is a significant technical challenge, involving mechanical construction, antenna work, low-noise receivers, high-power transmitters, tracking and astute operating.

EME used to be the exclusive preserve of those with very high power stations and large steerable antenna arrays; but the new weak-signal digital modes like JT65 mean that today even modestly equipped stations can experience EME contacts—especially if the station on the other side of the contact has high power and a large steerable antenna array! One South African has contacted over 100 countries on the 2 m band using EME, and several others are trying to emulate the achievement.

27.9 Amateur Satellites

There are a number of amateur satellites in orbit that will relay signals in various modes, including FM, SSB, CW and digital modes. They act like terrestrial repeaters, except that signals are usually sent up to the satellite on one band (the *uplink* band) and received from it on a different band (the *downlink* band). A pair of uplink and downlink bands is known as a *mode*. Modes are described using two letters, which represent the uplink and downlink bands, for example Mode V/U which has a 2 m uplink and a 70 cm downlink. The "V" stands for "VHF" referring to the 2 m band, and the "U" for "UHF" referring to the 70 cm band).

The easiest satellites to work are those in low earth orbit, which because they are fairly close to the earth (100 to 200 km above the surface) can be worked with low power and simple antennas. Because they are low they offer a fairly small footprint (the area in which stations can communicate via the satellite) and short pass times – often only a few minutes. The high earth orbit satellites like AO-40 offer a larger footprint and much longer pass times, but require more sophisticated equipment to access. Gain antennas may have to be steered in azimuth and elevation to track the satellite. Don't overdo it, as satellites have a limited power budget. If you put too loud a signal into the uplink receiver, the on-board AGC will reduce the receiver sensitivity, killing the signals from all other users. You will stand out like a sore thumb, and probably spend a lot of time answering fan mail.

Amateur satellites were traditionally given an Oscar number by Amsat, the international Amateur Satellite Corporation. OSCAR meant "Orbital Satellite Carrying Amateur Radio".

These days, numbers such as AO-40 mean "Amsat Oscar 40". AO-7 is the oldest surviving amateur satellite[9], dating from 1974.

Because satellites move at high speed, they exhibit Doppler shift. The operator may have to look for the downlink at a frequency slightly displaced from the nominal frequency, and may have to adjust the uplink frequency to a slightly different frequency to that used by the uplink receiver.

27.10 Propagation Prediction

Broadcasting, commercial and government users of radio have long had a need to predict propagation. As our understanding of the ionosphere has improved, the accuracy of forecasting has also improved.

Today, a number of free software tools can be used to predict propagation on any path, using solar data as input. Many of these tools are based on VOACAP, which was developed to forecast Voice of America coverage.

VOACAP is available with a bewildering array of user interfaces, some of which can draw fancy coverage maps showing expected signal strength over the entire globe using colour coding.

The output of a propagation forecasting algorithm is a *path loss*. This number (in dB) describes how much of the original signal will arrive at the destination. In principle, if you know how much power is being transmitted and how much will be required at the receiver, you can calculate the maximum path loss that can be tolerated, and the expected signal to noise ratio given the expected path loss.

For satellite and line-of-sight systems, the path loss can be readily calculated using the normal decay of signal strength with distance:

$$L_P = 20 \log_{10} (4 \pi d \div \lambda)$$

with L_P being the path loss (in dB) using isotropic antennas, d being the distance in m and λ being the transmission wavelength in m. In principle, any distance unit can be used, provided both quantities are expressed in the same units.

Using this calculation, the total link can be modelled, taking into account antenna gain, feeder losses, connector losses, polarisation mismatch losses and obstructions.

$$P_{Rx} = P_{Tx} + G_{Tx} - L_{Tx} - L_P - L_M + G_{Rx} - L_{Rx}$$

where

P_{Rx} is the received power [dBm]
P_{Tx} is the transmitter power [dBm]
G_{Tx} is the transmit antenna gain [dBi]
L_{Tx} is losses in the transmitter (coax, connectors etc.) [dB]
L_P is the path loss, calculated above or obtained from forecasting software [dB]
L_M is miscellaneous losses (polarisation mismatch, obstructions etc.) [dB]
G_{Rx} is the receive antenna gain [dBi]
L_{Rx} is losses in the receiver (coax, connectors etc.) [dB]

[9] If you exclude Oscar 0, the EME wisecracks' name for the moon…

Once the expected receive power is known, it can be compared with the receiver sensitivity specifications and the likely ambient noise conditions to determine the likely signal to noise ratio. This ratio (in dB) must be compared to the minimum required signal to noise ratio to determine the *link margin*, the amount of unanticipated loss that can be tolerated before the link will no longer function. A similar approach can also be followed to calculate the minimum required transmitter power.

Summary

Direct wave (line of sight) propagation is when signals of any frequency travel directly from the transmitter to the receiver. Ground-wave propagation is where low and medium frequency signals follow the curvature of the earth, up to a distance of several hundred kilometres.

Ionospheric propagation results from the refraction of radio waves by the E, F_1 and F_2 ionospheric layers. During daylight hours, the D layer absorbs low-frequency signals, so only higher frequencies are usable. The D layer dissipates rapidly after dark, allowing even low frequency signals to reach the F layer. Higher frequency signals are not refracted sufficiently by the ionosphere to return to earth, but are lost into space. The skip zone is the area within the first hop in which a signal cannot be heard, as it is too far for direct and ground wave, and too close for skip (perhaps 100 to 1000 km). The critical frequency is the highest frequency at which radiation directed vertically upwards will return to earth. The maximum usable frequency (MUF) for a particular path is the maximum frequency that will be refracted by the ionosphere along that path and it may be considerably higher than the critical frequency. The lowest usable frequency (LUF) is the lowest frequency that can be used for communication on a particular path, and depends on the EIRP of the transmitter and the receiver noise level as well as the extent of ionisation. Ionospheric propagation via the F layer occurs most commonly for the high frequency (HF) bands, although there are occasional openings on the 6 m band. The amount of ionisation depends on the time of day, season and the eleven-year solar cycle.

Multipath propagation could result in enhancement and destruction of the signal, leading to *fading*. Fading could vary the signal strength by tens of dB and could have a period of milliseconds to hours.

Sporadic E propagation consists of the refraction of VHF signals by intensely-ionised patches of the E layer. These patches occur sporadically but may last for several hours and allow VHF communication at ranges from a hundred to several thousand kilometres. Backscatter allows close-in contact on the high bands, when signals are scattered from the surface after the first hop. Meteor scatter mostly uses specialised digital modes to communicate using the very brief periods of intense ionisation caused by meteors entering the earth's atmosphere. Tropospheric scatter results from signals being reflected by temperature and humidity differences in the troposphere and can result in consistent VHF and UHF communications over ranges of 100 to more than 500 km with suitable equipment. Tropospheric ducting, when VHF signals are trapped between the ground and an inversion layer or between two inversion layers, is much less common but can result in signals being received with good strength thousands of kilometres away. Auroral scatter reflects rough signals from the polar regions, allowing long-distance VHF and UHF contacts.

Earth-moon-earth (EME) is technically challenging because of the extreme path loss. Nevertheless, EME is possible with relatively modest stations using weak signal digital modes.

Amateur satellites retransmit signals received on one frequency band onto another frequency band, functioning similarly to repeaters in space but over much greater distances

than terrestrial repeaters. Some satellites require high-gain antennas that have to track the satellite in azimuth and elevation. Operators may have to compensate uplink and downlink frequencies to compensate for Doppler shift. Do not use more power or antenna gain than necessary, to avoid inconveniencing other users.

Link budget calculations can be done by calculating free-space path loss or using a propagation forecasting tool. The calculations take into account transmitter power, antenna gain and losses, receiver sensitivity, antenna gain and losses and other factors like polarisation loss. The link margin provides an indication of how robust the link will be, what power is required, what antennas are required and what the reliability of the link is likely to be.

Revision Questions

1 What is the transmission path of a wave that travels directly from the transmitting antenna to the receiving antenna called?

a. The ground wave.
b. The sky wave.
c. The linear wave.
d. The plane wave.

2 What effect does tropospheric bending have on 2 m radio waves?

a. It increases the distance over which they can be transmitted.
b. It decreases the distance over which they can be transmitted.
c. It tends to garble phone transmissions.
d. It reverses the sideband of phone transmissions.

3 Two stations 5 km apart are most likely to be communicating via:

a. Tropospheric waves.
b. Ionospheric waves.
c. Direct waves.
d. Ground waves.

4 The D layer occurs in the ionosphere at a height of about:

a. 80 km
b. 150 km
c. 200 km
d. 300 km

5 The F_2 layer occurs at a height of:

a. 80 km
b. 150 km
c. 100 to 200 km
d. 200 to 300 km

6 The ionospheric layer that mostly reflects long distance radio communications is:

a. D layer.
b. E layer.
c. F_1 layer.
d. F_2 layer.

7 Signals above the MUF passing through the F_2 layer:
a. Are reflected to earth.
b. Pass through and are lost in space.
c. Are amplified.
d. Are attenuated and refracted.

8 A VHF station finds a propagation opening on 2 m that lasts for an hour, with contacts of around 1500 km. This opening is most likely caused by:
a. Sporadic E
b. Tropospheric scatter
c. Ionospheric refraction in the F layer
d. Meteor scatter

9 Meteor scatter QSOs:
a. Often use SSB.
b. Use very short pulses of propagation.
c. Are only possible in summer.
d. Are common on the lower HF bands.

10 EME communications are accessible to:
a. Superstations using massive antennas and high power.
b. All VHF stations.
c. Relatively modest stations using computer-based weak-signal modes.
d. *a* and *c*.

11 Satellite frequencies change while monitoring the satellite's signals during its pass. This change is due to the:
a. Height of the satellite.
b. Doppler frequency shift.
c. Drift.
d. The circular orbit shape.

12 Both Azimuth and Elevation refer to:
a. Satellite ground station antenna positions.
b. Mobile communications.
c. Maritime communication.
d. Doppler direction finding.

13 Satellites contain transponders which relay:
a. Only CW signals.
b. Only FM signals.
c. All modes of modulation.
d. Digital signals only.

14 When working via a satellite, you should:
a. Use the maximum power permissible.
b. Speak Esperanto.
c. Use sufficient power to maintain reliable communications.
d. Use a speech processor and shout for greater penetration.

15 A link budget is:

a. The amount of money needed to construct a communications system.

b. The cost of civil engineering required to establish antennas.

c. A calculation of the transmitter and receiver parameters, antennas and path loss to evaluate the feasibility of a communications system.

d. The business model for a commercial radio station.

Chapter 28: Electromagnetic Compatibility

28.1 Definition of Electromagnetic Compatibility

Electromagnetic compatibility (EMC) is the process of ensuring that equipment that radiates electromagnetic radiation, such as an amateur transmitter, does not interfere with equipment that may be sensitive to electromagnetic radiation, such as television and radio receivers, pacemakers and computers.

As more and more electrical gadgets come into operation, the problem of mutual interference becomes worse and worse. The electromagnetic spectrum is increasingly polluted to an extent that makes radio communications harder and harder. In large cities, the problem may become so bad that some have referred to *electromagnetic smog* as a way of describing the noise produced by millions of discrete devices in a city. EMC seeks to alleviate this pollution by reducing the amount of noise generated and by addressing the immunity of devices to that noise.

There are two considerations when dealing with interference problems. The first consideration is technical: the causes of interference, and how they can be eliminated. The harder consideration is a legal and social one: Who is responsible for solving the interference problem? The problem is both legal *and* social because often an attempt to use a legal approach will generate undesirable social results.

Interference can be caused by intentional or unintentional radiators, and can take place to a device that is a *receiver* (that is designed to receive radio signals) or is not a receiver (that is not intended to receive radio frequencies, but is experiencing interference nevertheless).

28.2 Intentional and Unintentional Radiators

An *intentional radiator* is a device that is intended by virtue of its function to radiate, such as an amateur transmitter or a garage opening remote control.

An *unintentional radiator* is a device that does not need to radiate in order to perform its intended function, such as a motor vehicle ignition system or an electric fence.

There are strict limits to the maximum permitted radiation from unintentional radiators. If a system that does not include a radio transmitter of some kind is causing interference, that is generally because the system is radiating more than permitted, and it should be repaired or replaced at the owner's expense.

For example, if you receive interference from a neighbour's electric fence, that probably indicates that the electric fence is radiating more than is permitted, and the neighbour is responsible for having the defect rectified, and must turn the electric fence off until it complies with requirements. Of course convincing your neighbour of this obligation may not be so easy!

In general, any switch generates some radio noise while switching high currents. As the contact is made or broken, some sparking may occur, leading to the transmission of a noise burst. This noise may interfere with radio communications at a considerable distance. In fact, millions of such switches in an urban environment contribute to a gradual raising of the noise level, until all radio communication within the city becomes hampered.

28.3 Interference to non-receiving equipment

The converse applies when the equipment being interfered with is not intended to receive radio signals. For example, suppose your neighbour reports that your radio transmissions are "breaking through" on their stereo system when they are listening to CDs. Because the

stereo system when listening to CDs is not supposed to receive radio signals, the problem lies with the stereo, not with the radio transmitter.

Often the root cause is that the affected equipment was not designed for, and has not been tested in, environments with strong RF signals present. Unfortunately it is quite legal for such equipment to be sold, and it will work fine for 99% of the time, since in most locations it will encounter only weak signals from distant transmitters. Then an amateur moves in next door, sets up equipment that is operating within the limits of their licence, and all of a sudden the neighbour's CD player receives interference. It is quite natural for the neighbour to think that this is the amateur's fault, and that they must fix the problem or stop transmitting. However, the fault lies with the manufacturer of the equipment for not designing it to withstand the levels of electromagnetic signals that may result from a nearby transmitter.

In this case, even though it is the neighbour's responsibility to solve the problem, it would be diplomatic for the amateur concerned to make his or her technical skills available to the neighbour to help diagnose the problem and suggest solutions. Apart from good neighbourliness, the same neighbour may have the opportunity to comment on your application to erect a tower, and is more likely to be kindly disposed to such a request if you have helped them to solve any problems that appear to have been caused by your transmissions in the past!

28.4 Intentional Radiators interfering with Receivers

The situation is slightly more complex if an intentional radiator (such as your amateur transmitter) interferes with a device that is intended to receive radio signals (such as your neighbour's television set). In this case, the key question is the nature of the interfering signal.

If the interfering signal is in all respects a legal licenced transmission—that is, it is within an amateur band, does not exceed the power permitted for the band and licence holder, and is a clean signal—then the problem is being caused by the receiving equipment being affected by an out of band signal, and it is the receiving equipment that is defective and must be repaired.

On the other hand, if the transmitted signal in any way does not conform with the requirements of your licence, you should first correct the problem with the transmitted signal before suggesting to your neighbour that they have their TV fixed! If interference is reported to ICASA (the regulator), their first course of action will probably be to inspect the transmitting equipment. If it is found to be out of order in any way, you may be held responsible for the interference and, even if you are not, the transmitting equipment can be confiscated if it does not comply with your licence requirements.

Once again, as a matter of diplomacy, it is a good idea to assist your neighbour if possible to solve the interference problem, even if you have determined that your transmitter is operating quite legally. As well as maintaining peace in the neighbourhood, this course of action will help to maintain the good reputation of amateur radio. However, if this is not possible—for example, if your neighbour refuses your assistance and insists that you just stop operating—then as long as you are certain that your equipment is operating legally, you are entitled to continue to operate despite the interference to your neighbour's television or other equipment.

When offering technical assistance to resolve interference problems, remember that you may be held liable for any changes you make to the neighbour's installation that may later lead to problems. If you solder a filter into your neighbour's speaker leads and the sound

system suddenly stops working a few months later, you are likely to have fingers pointed at you.

28.5 Shared Bands

One exception to our classification is that some amateur bands are shared between different users, with one of the users being declared the “primary” user and the other as “secondary” users. For example, amateur radio has been allocated the 13 cm band (2,3 to 2,45 GHz) on a secondary basis; the primary use is industrial, scientific and medical.

Simply put, secondary users may not cause interference to primary users (and must stop operating if this is the only way to prevent interference), while they must accept interference from primary users. So if you live next door to a hospital and receive interference from medical equipment that is intentionally radiating in the 2,4 GHz band, there is nothing you can do about it.

Of course all amateur bands are shared with other amateurs, and it is important that we take steps to avoid interfering with our fellow amateurs. These steps should include operating courtesy and ensuring that your transmitter is radiating a clean signal.

28.6 Causes of Interference

There are three possible causes of interference.

1. The transmitter may be radiating on a frequency that it should not be radiating on.

2. The receiver might be receiving signals that it should not be.

3. The transmitter and receiver may both be working correctly, but something else is translating the transmitted signal to the frequency of the receiver. For example, corrosion in a gutter can cause the metal to operate like a rectifier, re-radiating harmonics of signals transmitted from a nearby transmitter.

4. Other sources of noise, such as high-power electrical switches or motors, or even natural phenomena such as lightning, can cause noise that will influence receivers, both intentional and non-intentional.

Since the third mode is quite uncommon and usually requires specialised equipment and significant expertise to resolve, we will only look at the first two possibilities.

28.7 Transmitter Defects

The most common problems in transmitters are frequency instability, harmonic radiation, spurious oscillations, and “wide” signals.

Frequency instability is usually the result of LC (inductor/capacitor) oscillators that have not been adequately compensated for temperature variations or protected against mechanical shock. It is most likely to impact on other amateurs, unless the instability is sufficient to take the transmitter out of the amateur band and cause interference to other services. Fixing frequency instability usually requires design modifications or improved construction methods (for example, more solid construction that is less sensitive to mechanical knocks). It is quite uncommon with modern crystal-controlled synthesised radios, although it may occur if a PLL frequency synthesiser gets unlocked from the reference frequency.

Another type of frequency instability is chirp, which occurs when the oscillator frequency is affected by the loading of subsequent stages or by fluctuations in the power supply voltage

when a CW transmitter is keyed. It can be prevented by using a high-impedance buffer amplifier after the oscillator; and by regulating the oscillator voltage supply.

Harmonic radiation occurs on multiples of the transmitter output frequency. For example, a transmitter operating at 144 MHz may interfere with a television receiver operating at 720 MHz (144 MHz x 5). It can be caused by overdriving an amplifier stage (for example by having the microphone gain or CW drive level set too high) or by inadequate attenuation of harmonics by the transmitter's output lowpass filter (e.g. when the output controls on the amplifier are improperly adjusted).

If the problem is caused by overdriving the transmitter, the solution is to reduce the drive level by adjusting the microphone gain or CW drive correctly. However, if the problem persists even when the transmitter is not being overdriven, the best solution is to add an additional lowpass filter between the transmitter and the antenna. Lowpass filters for the HF bands (up to 30 MHz) are available at reasonable cost and provide substantial attenuation at higher frequencies, typically 50 dB or better at 50 MHz.

Another solution sometimes recommended is to use an antenna tuning unit (ATU) even when it is not required to match the antenna, as the ATU may attenuate out-of-band signals. When doing so, ensure that the ATU is a lowpass filter, and not a highpass filter. A Pi network with a single inductor or a T network with two inductors should do the trick.

Spurious oscillations may either be self-oscillation, at or near the intended frequency of operation of an amplifier or mixer, or parasitic oscillations, which usually occur at VHF or UHF frequencies far away from the intended frequency of operation. Self-oscillation is caused by unintended feedback from the output of an amplifier or mixer that includes tuned circuits to its input, causing oscillation at the resonant frequency of the tuned circuit. It can be suppressed either by reducing the coupling (for example by shortening component leads) or by introducing negative feedback to reduce the loop gain and prevent oscillation.

Parasitics are VHF or UHF oscillations that occur due to unwanted "hidden" resonances in oscillators and amplifiers—for example, between RF chokes and decoupling capacitors, or due to the inductance of capacitor leads at high frequencies. They can be eliminated by using low-Q (lossy) RF chokes, which are less likely to cause oscillations, or by using ferrite beads to add sufficient inductance to component leads or wires to dampen out unwanted VHF or UHF oscillations.

"Wide" signals are signals where the bandwidth exceeds the minimum required due to inter-modulation distortion. The cause is usually that some amplifier stage is being overdriven, and while this may result from a design defect it is more often caused by an incorrectly adjusted microphone gain control or CW drive level. On most modern transmitters the ALC (automatic level control) voltage can be monitored on the transmitter's meter during transmissions. The microphone gain or CW drive level should always be adjusted so the voltage remains within the acceptable ALC levels at all times. These levels are usually marked on the meter.

Another cause of wide signals is amateurs intentionally "opening up" the audio paths on their transmitters to allow the broadcast of wideband audio signals that exceed the 3 kHz bandwidth required for communications quality in the pursuit of "fidelity", but at the cost of causing interference to other operators.

A CW transmitter may generate key clicks if the carrier is switched on or off too rapidly when keying. The carrier should be turned on or off gently over a period of about 5 ms to avoid generating key clicks. Unfortunately, even some very well-regarded modern

transceivers like the original FT1000 MP have a problem with key clicks and may need to be modified to reduce clicks to acceptable levels.

If a high-power stage is keyed directly, arching of the key contacts may result. The solution is to key a lower-power stage and then feed the resulting keyed signal to amplifier stages, or to use an intermediate switch like a transistor or relay to do the actual keying, keeping the high power away from the key itself.

Mains hum may be heard on transmitted signals if the power supply is inadequately filtered. The addition of a voltage regulator or additional smoothing capacitors should solve the problem.

If a transmitter is using an antenna like a long wire that is driven against earth, it is important to have a good RF earth system that is independent of the mains earth. The earth lead must be as short as possible and must be routed as directly as possible. The mains earth wire usually travels in close proximity to the other mains wires for some distance before being physically earthed, so RF signals in the mains earth are likely to be inductively coupled to the live and neutral wires and may travel through them to neighbouring buildings, causing interference, especially to mains-operated equipment. The mains earth also often has high impedance at RF frequencies, so an independent earth system is necessary to remove RF voltages from equipment and antenna feed-lines. Of course, even if you cannot provide a good RF earth, a mains ground is still required to prevent the case from having a potentially lethal voltage in the case of a fault.

28.8 Receiver Defects

The most common defect in radio and television receivers that results in interference from amateur transmissions is *receiver overload*. Signals stronger that the receiver was designed to handle are present at the receiver input, and inter-modulation distortion in the first mixer causes spurious products that interfere with reception.

One common cause of over is inexpensive masthead RF preamplifiers that are sometimes used to improve television reception in marginal areas. While preamplifiers with decent signal-handling capabilities are available, they are generally more expensive, and the inexpensive ones that are widely available are very prone to overloading.

A solution to receiver overload is to add additional filtering before the receiver that removes the strong out-of-band signals that are overloading the receiver. What type of filter is required will depend on what frequency transmissions are causing interference. If transmissions in the HF bands are causing the problem, a highpass filter between the TV antenna and the TV might solve the problem, since the TV transmissions are on higher frequencies in the VHF and UHF region, so these frequencies can be passed while blocking HF frequencies. Obviously, if a masthead preamplifier is in use, the filter must be on the mast too, between the antenna and the preamplifier.

If amateur VHF transmissions are interfering with UHF television reception, a highpass filter with a cutoff frequency of 470 MHz might solve the problem. However, if VHF transmissions are interfering with VHF television reception, a bandstop filter for the particular interfering amateur transmission band might be required. These bandstop filters are also called “traps”. A quarter-wavelength transmission-line “stub” connected across the feed-line and open at the far end, may also serve as a trap. It presents a low impedance at the frequency on which it is exactly a quarter wavelength, effectively shorting the two conductors in the feed-line together at that frequency, while presenting a high impedance at most other frequencies. Signals move more slowly in coaxial cable than in free space, so a quarter-wavelength stub is shorter than an actual quarter wavelength.

However, note that if the problem is being caused by overloading a masthead RF preamplifier, no amount of filtering of the signal between the amplifier and the television will help, as in-band spurious products may already have been generated by the amplifier. In this case, replacing the amplifier with one that is more resistant to overload (or removing it altogether if reception conditions permit) may be the only option.

Interference to receivers may also result from *image signals*, also known as *second-channel* interference, if the image frequency of a receiver coincides with the frequency on which a strong amateur signal is present and the receiver has insufficient image rejection.

Assessing Interference Sources

When hearing interference from a nearby transmitter, the operator must decide whether the interference is caused by the transmitter or by the receiver. Because transmitters and receivers use very similar techniques to generate and demodulate the signal, they suffer from very similar types of interference.

The key in determining whether it is a transmitter or a receiver problem is the fact that most interference is caused by saturation of some kind. If the transmitter is driven too hard, the signal will be distorted because the final amplifier cannot handle the signal amplitude, resulting in adjacent-channel interference. Likewise, a receiver that is overloaded with a very loud nearby signal will saturate, causing very similar interference.

The key is in reducing the incoming signal strength in the receiver. Most communications-grade receivers include a switchable attenuator for exactly this purpose. In other cases, receivers feature a preamplifier that is used for weak-signal work, which can be turned off to reduce signal levels. Finally, the receiver's RF gain can be reduced. Either way, the reduced signal strength will probably solve the problem if it is receiver-generated.

When hearing a "wide" signal, simply engage the attenuator and observe the effects. Most attenuators provide something like 20 dB of attenuation. If the adjacent interference disappears or decreases by 40 dB or more when the attenuator is engaged, the problem is in the receiver. However, if the adjacent interference decreases by only 20 dB when the attenuator is engaged, the problem is in the transmitter. A friendly request to the offending operator may be in order.

Remember that any of these measures to reduce adjacent interference will also reduce the sensitivity of the receiver. When you insert attenuation, turn off the preamplifier or reduce the RF gain, you are sacrificing sensitivity in exchange for a reduction in interference. Even though the incoming signal becomes weaker, the resulting SNR has improved.

28.9 Common-Mode Chokes

Interference usually "gets into" the equipment being interfered with through the wires attached to it. These wires include antennas, speaker leads, interconnections between audio components and mains power leads. In common-mode interference, the signal is transmitted in phase by both the conductors in the connection—for example by both the live and neutral wires in the mains, or both conductors in the speaker cable, or both the inner conductor and the earth in a coax cable.

Common-mode interference can be effectively eliminated by a common-mode choke, also known as a "*braid breaker*". Although it does not involve physically breaking the braid in a coax cable, it effectively blocks the flow of common-mode signals that travel along the braid as well as in the inner conductor, which is where the name comes from.

The choke consists of several turns of the cable—which could be a mains power lead, a speaker cable, or a coax cable—wound around a suitable core to form an inductor. Ferrite

toroidal cores are the best, and are available for the purpose from local suppliers. The idea is that common-mode currents will generate a magnetic field in the core, and so the choke will act as an inductor to common-mode signals. If the inductor has sufficiently high impedance at the frequency causing the interference, this signal can be rejected.

However, differential signals—that is, signals where currents flow in opposite directions in the two conductors, for example the signal from the antenna in a TV antenna lead—will not generate a magnetic field since the fields generated by the two currents flowing in opposite directions cancel out; and so the common-mode choke does not act as an inductor for differential signals, which pass through unaffected.

Common-mode chokes can be used both with receiving equipment, such as television receivers, and with non-receiving equipment such as audio amplifiers that are suffering interference from strong radio signals.

28.10 Direct Radiation and Shielding

In a few cases, electronics may be directly influenced by strong electric fields. Currents can be induced into circuits without going through connecting cables.

The problem normally manifests when very strong electric and magnetic fields are present. The most likely situation is for equipment situated near an antenna connected to a high-power transmitter. In such cases, interference can be coupled directly into the IF stages. Such IF interference is characterised by it being present on all channels or frequencies that the device is capable of receiving.

Remember that an electromagnetic signal is composed of E and H fields. The E field is measured in V/m and the H field in A/m. These fields can be measured using a field strength meter.

The problem is normally solved by good design practices—using decoupling components such as capacitors within the circuit itself—and by *shielding*. Consumer devices are normally made very cheaply, and often do not comply with good design practices. Manufacturers have a duty to solve problems that are due to design inadequacies, but local distributors are not always willing and capable to do so. The problem is therefore more likely to be solved by good shielding.

Shielding consists of conductive enclosures that completely surround the circuitry, known as a *Faraday Cage*. Shielding should be solid or have only small holes. Holes that have a circumference of more than a fraction of a wavelength of the offending signal will be penetrated by the signal, leaving the equipment vulnerable to the strong field. Shielding also serves to reduce radiation of objectionable interference by the electronic devices.

Shielding against electric fields is relatively easy, as most metals are conductive to electricity and can be used to make Faraday Cage enclosures. If the coupling mechanism is predominantly magnetic, the problem is much harder to solve. Specific materials such as Permalloy or Mu Metal must be used, and the shielding must be relatively thick. Magnetic fields are normally not the dominant problem at higher frequencies (HF, VHF and up).

Transmitters can also be prone to direct radiation, this time outbound. Some transmitters can radiate energy that does not go through the antenna connector, through so-called *cabinet radiations*. The cause is normally stray currents inside the cabinet. As with susceptibility problems, the solution is not easy, as the bad equipment design is probably not easy to fix. The problem must be solved by improving screening. The transmitter must be inside a Faraday Cage. If the existing enclosure is not good enough, more work may be required, analogous to the suggestions for susceptibility given above.

28.11 Sensible Measures against Interference

Many types of interference can be alleviated by simple courtesy. Mount your antennas as high and as far from potential interference as you can. Use the minimum power required to facilitate the communications of the moment. Listen before you transmit. Much of the interference that results in practice is due to a violation of one or more of these simple rules.

Summary

EMC should be looked at from two perspectives: the technical (*how* to solve the problem) and the legal (*who* is responsible for solving the problem). If the interfering signal is being generated by equipment that does not need to transmit in order to function, it is this *unintentional radiator* that is usually at fault since there are strict limits as to how much electromagnetic energy can be radiated by unintentional radiators. If the equipment being affected is not intended to receive radio signals of some kind, the affected equipment is at fault. If a signal from an intentional radiator is affecting equipment that is designed to receive radio signals, the key question is whether the transmitter is operating within the frequency and power limits specified by its licence. If the transmitter is not radiating legally, the exceedances must be fixed. However, if the transmitter is operating correctly and within licence requirements, the problem is being caused by the affected equipment responding to an out-of-band signal, and ultimately it is up to the owner of the affected equipment to have the problem repaired at his or her expense.

However, it is advisable for an amateur whose transmissions are causing interference to assist as much as possible in diagnosing the cause of the problem and suggesting solutions. This is both to maintain a good relation with neighbours and to maintain the good image of amateur radio. Just be wary of making changes to the neighbour's installation, as subsequent problems with the equipment may well be blamed on the helpful radio amateur.

The most common transmitter problems are frequency instability, harmonic radiation, "wide" signals and key clicks. Frequency instability requires due attention in design and construction to temperature compensation, mechanical rigidity and suitable buffering of oscillators to avoid chirp. Harmonic radiation can be attenuated by a suitable lowpass filter. Wide signals are usually caused by setting the microphone gain level too high. Key clicks are the result of turning the carrier on or off too rapidly.

Receiver problems can be caused by common-mode or differential signals. Common-mode signals can be attenuated by a suitable common-mode choke (also called a "braid breaker"). Differential-mode signals require the use of suitable highpass or bandstop filters between the antenna and the receiver. Mast-head TV amplifiers are often subject to overloading. The amplifier may need to be removed or replaced with one that is less subject to overloading.

An attenuator can help to diagnose interference being received. If the attenuator attenuates the interference just as much as the signal causing it, the problem is in the transmitter. If the attenuator completely cures the interference or reduces it by much more than the offending signal, the problem is in the receiver.

Strong electromagnetic fields can couple directly into electronic equipment. The solution is good design of the target electronics and thorough shielding (a *Faraday Cage*). Shielding should not have large holes, failing which the radio signals will still penetrate the enclosure. For coupling that is predominantly magnetic, special enclosures of special materials will be required.

Transmitters can also suffer from cabinet radiations. Fixing these problems is similar to the suggestions for shielding given above.

Simple courtesy requires that you operate in a way that minimises the risk of interference. Very often, that's all that is required.

Revision Questions

1 EMC defines the compatibility of electronic equipment to:

a. Static noise.
b. Man made electromagnetic noise.
c. High supply voltages.
d. Battery operated equipment.

2 One aim of EMC is to:

a. Prevent pollution of the RF spectrum.
b. Encourage high power transmissions.
c. Discourage development of amateur radio.
d. Desensitise radio receivers.

3 Spurious oscillations caused by resonant of RF chokes can be minimised by using:

a. Low Q chokes.
b. Long power cables.
c. Non-inductive capacitors.
d. Non-resonant circuits.

4 Self oscillations can occur when the output of an amplifier is coupled to:

a. An antenna.
b. A dummy load.
c. A pi-filter network.
d. The amplifier input.

5 An RF power amplifier is found to oscillate at its fundamental frequency when the RF drive is removed. This effect is called:

a. Self-oscillation.
b. Parasitic oscillation.
c. Harmonic oscillation.
d. Overload oscillation.

6 The cure for-self oscillation in an audio amplifier is:

a. To increase voltage gain.
b. To filter the feedback signal.
c. To inductively couple the input stage.
d. To introduce negative feedback.

7 Insufficient carrier suppresion on an SSB signal will cause:

a. distortion.
b. poor readability.
c. difficulty to set the receiver BFO.
d. heterodynes on the audio frequencies.

8 To minimise mains hum on transmitted signals, all DC power supplies should:

a. Use a low DC voltage.
b. Use a screened transformer.
c. Be RF decoupled.
d. Use smoothing and regulator circuits.

9 A 1000 µF capacitor across the DC output of a power supply:

a. Will increase any 100Hz ripple present.
b. Improve low frequency response.
c. Remove AC rectified mains hum.
d. Decrease smoothed output voltage.

10 To minimise interference on adjacent channels, voice frequencies should be kept below:

a. 500 Hz
b. 1 kHz
c. 3 kHz
d. 5 kHz

11 So as not to cause unnecessary sideband splatter, the percentage modulation of an AM signal must be kept below:

a. 25%
b. 50%
c. 75%
d. 100%

12 What causes splatter?

a. Inadequate harmonic suppression in the final amplifier.
b. Excessive bandwidth of a transmitter.
c. A poorly regulated transmitter power supply.
d. Insufficient drive to the final amplifier.

13 Intermodulation caused by a linear SSB amplifier is due to:

a. Over driving the power level of the amplifier.
b. The operating frequency being to high.
c. Harmonic distortion.
d. Two modulating frequencies occuring at the same time.

14 Over-driving an SSB linear amplifier can cause:

a. Improved communication.
b. A louder audio signal.
c. Lower power consumption.
d. Distortion and splatter.

15 Which of the following might be effective at reducing the risk of parasitic oscillations in a low power VHF output stage?

a. Ferrite beads on the emitter lead of the power device.
b. Ferrite beads on the microphone cable.
c. Ferrite beads in series with the microphone.
d. Ferrite beads on the loudspeaker leads.

16 Parasitic oscillations can cause interference. They are:

a. At a very low frequency.
b. Always at twice the operating frequency.
c. High in frequency but not related to the operating frequency.
d. Always at three times the operating frequency.

17 Any non-linear device will produce:
a. Mixing products.
b. Amplification.
c. Filtering.
d. Key clicks.

18 When a synthesised VFO oscillator is not locked to the reference frequency, it will be:
a. Stable.
b. Equal to the reference frequency.
c. Unstable.
d. Equal to the operating frequency.

19 A domestic receiver having an IF of 455 kHz and receiving a signal on 945 kHz, experiences strong breakthrough from someone on the 160 m band. This interference could be caused by second channel interference of:
a. 1,810 MHz.
b. 1,825 MHz.
c. 1,835 MHz.
d. 1,855 MHz.

20 A typical source of polluting electromagnetic interference is caused by:
a. Electric musical instruments.
b. Video signals.
c. Audio signals.
d. Arcing electrical switches.

21 A lowpass filter is most likely to be found in:
a. A crystal oscillator.
b. The output stage of an HF transmitter.
c. A TV antenna amplifier.
d. A mixer.

22 A ferrite bead around a piece of wire:
a. Decreases the wires impedance.
b. Protects the wire from damage.
c. Blocks the flow of RF signals along the wire.
d. Improves power dissipation.

23 A braid breaking toroidal choke wound onto a coax feedline:
a. Passes anti-phase currents.
b. Blocks anti-phase currents.
c. Passes in-phase common mode noise.
d. Acts as a balun.

24 An interfering signal picked up by a long feedline can be attenuated by:
a. Raising the receiving antenna.
b. Replacing the feedline.
c. Correctly matching the feedline.
d. Installing a toroidal choke.

25 In RF power amplifiers, the DC wiring associated with the tank circuit often passes through ferrite beads. The beads:
a. Introduce local lowpass filters in the wiring.
b. Cause high power losses at VHF.
c. Act as fine tuning controls for the tank circuit.
d. Increase the "Q" of the tank circuit.

26 To eliminate RF pickup on the outer screen of a coax cable:
a. Install a balun.
b. Remove the earth from the coax cable.
c. Install a braid breaker.
d. Use lower loss coax cable.

27 A TV antenna coax feedline picks up an amateur transmission. This interference can be resolved by trying to install:
a. A masthead amplifier to override the incoming interference.
b. A braid breaker.
c. New TV coax cable.
d. Filters on the mains power plugs.

28 It is found that interfering signals are being induced on the braid of an antenna downlead to a domestic FM radio by a 144 MHz transmitter. One possible solution is:
a. To fit a braid breaker filter on the antenna feedline.
b. Remove the 144 MHz transmitter earth lead.
c. To increase the 144 MHz transmitter power.
d. To fit the 144 MHz transmitter with a lowpass filter.

29 The antenna of an amateur station must be located in a position that:
a. Is easily accessible.
b. Is in line with other power lines.
c. Will not induce high field strengths in domestic premises.
d. Is below all other structures.

30 The location of the feeder of an amateur antenna must be
a. Of a precise length.
b. Kept away from other cable routes.
c. Not visible.
d. Kept close to other telephone cables.

31 The earthing of an amateur station is required to:
a. Give the mains a good earth.
b. Minimise undesired RF voltages on the feeder and equipment.
c. To prevent mains earth leakage.
d. Enable the equipment to operate from batteries.

32 When operating a mobile HF set at home from a battery supply using the base antenna, there is no breakthrough problem. When using the same arrangement with an earthed battery charger also connected, breakthrough occurs on an electronic organ. The possible cause is:
a. the production of harmonics at the transmitter.
b. very strong received signals.
c. Poor RF earthing.
d. RF earthing is too good.

33 To minimise harmonic radiation most HF transmitters contain:
a. A highpass filter.
b. A notch filter.
c. A lowpass filter.
d. Bandpass filters.

34 The term "trap" when discussing filters describes a device which:
a. Increases signal output.
b. Narrows the bandwidth of an antenna.
c. Acts as a notch filter.
d. Acts as a dummy load.

35 The length of a coaxial trap used to filter out an interfering signal is:
a. A quarter wave length of the interfering signal.
b. A random length.
c. The wavelength of the transmitted signal
d. 250 mm

36 A notch filter one quarter wavelength long used to filter out an interfering signal on the VHF bands is called:
a. a stub.
b. a balun.
c. a transformer.
d. an antenna tuning unit.

37 The main reason for providing substantial mains earthing points on RF equipment is:
a. To provide a path for RF to be bypassed to earth.
b. To provide a path for fault currents to be passed to earth.
c. To bypass all spurious signals to earth.
d. To increase earth resistance.

38 The leads used to connect RF equipment to earth should be:
a. Connected to the nearest mains plug earth terminal.
b. As short as possible.
c. Bare copper wire.
d. Connected via a suitable resistor.

39 In order to prevent the feeder to an antenna from radiating it should be:
a. As long as possible.
b. Cut to an exact length.
c. Screened and earthed.
d. Run close to the antenna.

40 In considering the equipment and power levels in a densely populated neighbourhood, it might be advisable to:
a. Keep the antenna as low as possible.
b. Locate the antenna as remotely as possible from the neighbours.
c. Use maximum output power.
d. Always use long feedlines.

41 The best place for an HF beam to minimise interference for an amateur living in a semi-detached house is:

a. On the joint chimney stack in the centre of the roof.
b. Overhanging the next door's roof space.
c. As high and far away as possible.
d. As low and far away as possible.

42 If interference is being coupled directly into electronic equipment by a nearby antenna, the best solution is probably to:

a. Add chokes to all cables connected to the equipment.
b. Enclose the equipment in a Faraday cage.
c. Enclose the equipment in Mu Metal or Permalloy.
d. Change the design of the electronics.

Chapter 29: Measurements

Measurements are important to determine whether equipment is functioning properly and to diagnose faults. This chapter introduces some of the measurements of interest to amateurs and the test equipment we use to make these measurements.

29.1 The Ammeter

The ammeter is used to measure current. In its simplest form, it consists of a coil through which the current to be measured flows, mounted on a bearing and suspended between the poles of a magnet. A current flowing through the coil generates a magnetic field, which will interact with the magnetic field from the permanent magnet, causing the coil to pivot on its bearings. This rotation moves a pointer attached to the coil, which indicates the current flowing on the meter scale. This mechanism is called a *moving-coil meter*.

An ammeter is connected in series with the wire in which the current to be measured is flowing, so that the current flowing through the wire also flows through the ammeter. In order to have the least effect on the circuit under test, the ammeter should have as small a resistance as possible.

The range of an ammeter can be extended by connecting a resistor, called a *shunt*, in parallel with the ammeter. The purpose of the shunt is to cause only a small part of the current being measured to flow through the meter, allowing the meter to measure a larger current than it was originally designed to. The shunt resistance can be calculated using the formula:

$$R_S = R_M / (n - 1)$$

where R_S is the shunt resistance, R_M the resistance of the ammeter, and n is the scale factor—that is, the ratio between the desired full-scale meter reading and the full-scale reading of the meter without a shunt. For example, suppose you want to measure a current of up to 1 A using a meter with a full-scale deflection of 1 mA and an internal resistance of 100 Ω. The scale factor is 1000 (to increase the full-scale deflection current from 1 mA to 1 A), so

$$\begin{aligned} R_S &= 100\ \Omega / (1000 - 1) \\ &= 0{,}1001\ \Omega \end{aligned}$$

Ammeters designed for small currents are generally called *milliammeters* or *microammeters*.

29.2 The Voltmeter

Voltmeters are used to measure voltage. A milliammeter can be converted into a voltmeter by adding a suitable *multiplier* resistor in series with the milliammeter. For example, suppose a milliammeter with a full-scale deflection of 100 μA and an internal resistance of 1 kΩ is required to measure voltages up to 10 V. The total resistance of the milliammeter plus the multiplier can be found by applying Ohm's Law:

$$\begin{aligned} R &= V / I \\ &= 10\ V / 100\ \mu A \\ &= 100\ k\Omega \end{aligned}$$

Since the internal resistance of the milliammeter is 1 kΩ, the series resistor required is 99 kΩ.

A voltmeter is connected in parallel with the component across which the voltage is to be read. In order for it to have the least effect on the circuit, the resistance of a voltmeter should be as high as possible. Transistorised voltmeters, using transistors, FETs or other devices to buffer the input can have in input resistance of many mega-ohms.

Moving coil meters are usually designed to measure DC. In order to measure AC voltages, a simple rectifier circuit may be employed. This simple rectification results in the meter measuring the *average* value of the rectified AC waveform, not the RMS value. However, the meter scales for AC voltmeters are usually calibrated so that if the waveform is a pure sine signal, the scale will read the RMS value. However, for waveforms other than sine signals, the reading will not be an accurate RMS value.

29.3 The Multimeter

The multimeter is a common piece of test equipment that uses a moving-coil or digital meter to measure voltage, current and resistance. Some multimeters may also measure capacitance and other quantities. The user must select the function (resistance, current or voltage) using a mode selector switch. In older meters, the user also has to select the scale (i.e. the full scale reading), failing which the meter may show too little deflection to read be readable, or the meter may be damaged by an overload. More modern meters automatically select the appropriate scale.

29.4 Frequency Counter

The frequency counter consists of digital circuits that count the number of cycles of the input waveform in a certain period, and then use this number to calculate and display the frequency of the input signal on a digital display. The accuracy of a frequency counter depends largely on the accuracy of the internal reference oscillator used to time the counting period. Using a crystal oscillator, the frequency counter may be very accurate, but if it is a simple inductor/capacitor oscillator the frequency counter may have an error of several percent.

29.5 Power and SWR Meter

Power meters measure the power output of a transmitter. The meter generally has input and output connectors and is installed in the feedline between the transmitter and the antenna.

Depending on the meter, it may measure the *average* power or the *peak* power. The distinction is especially important for phone signals as the human voice has much higher peak amplitude than average amplitude, and this difference will be reflected in AM and SSB signals. FM signals have constant transmitter output, irrespective of the amplitude of the modulating signal. Power meters are sometimes called *wattmeters*.

SWR meters generally measure both forward and reflected power, and use the ratio between forward and reflected power to calculate the SWR. Because they measure reflected power, they are sometimes called *reflectometers*. The term is not preferred, as it also describes measuring tools in several other fields, including optics.

Some modern SWR meters, called *antenna analysers*, include a built-in low power variable frequency oscillator and a frequency counter. These analysers makes it easy to measure the SWR at the antenna (as opposed to at the transmitter end of the feed line) and also allows measurements to be taken outside the amateur bands, as the built-in oscillator is so low-powered that it is legal for use on frequencies not allocated to amateurs. These meters suffer from one shortcoming, though: They tend to work badly in the presence of strong RF fields, as they interpret these signals coming down the feedline as reflected power. In these situations, higher power must be used to measure SWR, or a selective voltmeter that will reject the surrounding signals must be used.

29.6 The Oscilloscope

An oscilloscope displays signals that change rapidly, as mechanical meters cannot respond quickly enough to changes in the signal. It is ideal for measuring RF waveforms, modulation and other operating parameters in a radio station.

Old-style oscilloscopes contained a *cathode ray tube* (CRT) that displayed a dot on the display. The position of the dot could be adjusted from left to right by the voltage applied to the *X deflector plates* and up and down by the voltage applied to the *Y deflector plates*. The X deflector plates were usually driven by a time-base that generated a smoothly increasing voltage, causing the dot to sweep from left to right in a period set by the user, and then to return very rapidly to the left-hand side again before starting another sweep from left to right. The Y deflector plates were driven by the input voltage, usually through an amplifier (called the *Y amplifier*), causing the dot to deflect up or down according to the input voltage.

Modern oscilloscopes use a microprocessor to digitise the incoming signal and then display the same information on a screen, or directly on the user's computer or mobile device. The user interface is typically very similar to that of the old-style oscilloscopes.

The oscilloscope displays a graph of voltage (on the Y axis) against time (on the X axis) on its screen. The time-base is synchronised by a *trigger* circuit that starts the sweep from left to right when the input reaches a certain voltage. This synchronisation means that if the input consists of a repeating waveform, the display will "stand still" on the oscilloscope screen as each successive cycle of the input waveform traces the same pattern on the cathode ray tube display. The X-axis is usually calibrated in s/div (seconds per block on the on-screen graticule) and the Y-axis in V/div (volts per block on the on-screen graticule). The user can usually select scales of perhaps 100 ns/div to 10 s/div and from 1 mV/div to 100 V/div.

29.7 Marker Generator

A marker generator is a piece of test equipment that was used to determine the frequency of a receiver before frequency counters were available. It consists of a crystal oscillator that has been designed to generate harmonics that serve as frequency "markers" throughout the HF spectrum. A specific marker generator might be able to generate harmonics every 1 MHz, 100 kHz or 10 kHz depending on a switch setting. While turning the tuning dial, the user could then find the nearest 1 MHz marker, count the number of 100 kHz markers from there, and then the number of 10 kHz markers from there to get an accurate frequency measurement. Almost all modern transceivers include accurate digital frequency readouts, so marker generators are rapidly becoming obsolete.

29.8 The Dip Meter

The dip meter is used to measure the resonant frequency of a tuned circuit or antenna system. It consists of a variable frequency inductor/capacitor oscillator that is laid out so that the oscillator coil is accessible (usually plugged into a socket on the outside of the dip meter) and can be brought near to the tuned circuit being tested. The frequency of the oscillator is then varied, and as the frequency approaches the resonant frequency of the tuned circuit, energy is coupled from the oscillator coil to the tuned circuit and a "dip" is noted on the meter. The device is often called a grid-dip oscillator (GDO), from the time when vacuum tubes were used to observe a dip in the grid current during measurements.

29.9 The Dummy Load

A dummy load consists of a non-inductive resistor (usually 50 Ω) with sufficient power handling capability to dissipate the output of a transmitter being tested. It allows transmitter tests to be carried out without actually transmitting a signal. Transmitting a signal during testing when not strictly necessary would waste bandwidth and is poor operating practice.

Be careful if you build a dummy load since most high power resistors are wire-wound. These have considerable inductance and are not suitable for RF use. Practical dummy loads contain big cooling fins, fans and oil baths, or a combination of these, to get rid of the heat generate during testing.

29.10 The Field Strength Meter

The field strength meter consists of a small antenna, a diode detector and a sensitive microammeter. It is used to measure the strength of radio signals, for example to determine the directivity and approximate gain of an antenna. Simple field strength meters are generally not frequency selective and will respond to the presence of RF energy over a wide range of frequencies. They are also generally not very sensitive, requiring strong fields to produce a reading.

29.11 The Absorption Wavemeter

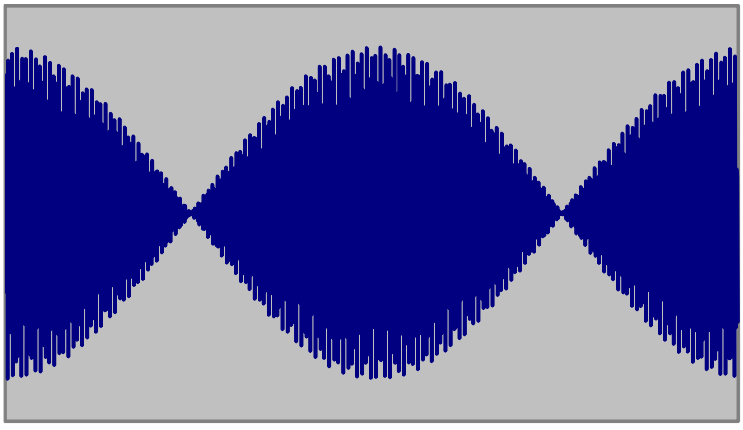
The absorption wavemeter is essentially a frequency selective field strength meter. It consists of an antenna, a tuned circuit to select the frequency, a diode detector and a microammeter. The purpose is to detect RF emissions on particular frequency bands. Because the tuned circuit is usually not very selective, it cannot be used to identify the precise frequency of a signal, but can be used to determine the approximate frequency. It is especially useful for detecting any harmonic radiation from a transmitter. For example, if you are operating a transmitter in the 80 m band but detect energy in the region of 7 MHz, it is a good indication that your transmitter is radiating harmonics.

Accurate measurements can be made with high-class calibrated measuring receivers with well-calibrated antennas. In general, to be meaningful such measurements must be made in laboratory conditions, such as anechoic chambers.

29.12 The Two-Tone Signal Generator

A two-tone signal generator generates an audio test signal consisting of two tones of equal amplitude that are not harmonically related. This signal is applied to the microphone input of an SSB transmitter in order to test it for linearity and to determine the peak envelope power if a peak-reading wattmeter is not available. The output of the transmitter is connected to a dummy load, and an oscilloscope is used monitor the waveform across the dummy load.

The following graph shows what the output of an SSB transmitter looks like on an oscilloscope when its input is connected to a two-tone test generator and it is operating linearly.



Output of an SSB transmitter with a two-tone test signal as input

If the output of the transmitter when viewed on an oscilloscope does not look like this, it is not operating linearly. Specific problems include “flat-topping”, when the curved tops and bottoms of the test signal are chopped off, which indicates that the transmitter is being overdriven. If successive cycles of the test signal do not join smoothly, but rather have a gap in between, it indicates that the amplifier is incorrectly biased. Either problem will result in inter-modulation distortion and must be fixed before the transmitter is used on air.

A more accurate measure of linearity may be obtained by viewing the output of the transmitter using a *spectrum analyzer*, which breaks the signal down into its component frequencies and plots the relative amplitude of the various components against frequency.

A two-tone generator test signal and an oscilloscope can be used to measure the peak envelope power of an SSB transmission if a peak-reading wattmeter is not available. It is impossible to calculate the peak envelope power of a signal modulated by a human voice from the average power, since the peak to average power ratio differs considerably between different voices and at different times. However, the peak to average ratio of the two-tone test signal is precisely 2:1. If you want to set an amplifier for a maximum of 400 W PEP (peak envelope power), you can apply a two-tone signal and adjust the amplifier until the average power output read on a wattmeter is 200 W. The peak power is then 400 W. An oscilloscope can be used to observe the amplitude of the modulation peaks. If voice modulation is applied, as long as the peak output is kept at this level, the PEP will still be 400 W.

Summary

An ammeter measures current. It is connected in series with the test circuit and must ideally have very low internal resistance. An ammeter can be made to measure more current than it was designed for with a shunt resistance.

A voltmeter reads voltage. It is connected in parallel with the circuit element to be tested, and must ideally have a very high input resistance. A voltmeter can be scaled to read higher voltages than its design capacity with a series resistor.

Multimeters can measure voltage, current and resistance, and possibly other quantities too. The user has to select the mode and possibly also the scale.

Frequency counters count pulses in a circuit and display a frequency in Hz. With an accurate timebase, such counters can be very accurate.

Power and SWR meters are installed in the feedline between the transmitter and the antenna. They can measure forward and reflected power and also display the SWR.

An oscilloscope can display rapidly-changing signals that change too quickly for mechanical meters. They normally display a timebase graph of the signal, with the time on the horizontal scale and the voltage on the vertical scale.

Marker generators emit signals rich in harmonics, to allow calibration of a receiver. With frequency counters now freely available, they are fading into oblivion.

A dip meter is an oscillator with an exposed coil, and can be used to determine resonance in tuned circuits and antennas.

A field strength meter can measure RF field strength. It is normally not very sensitive or selective. An absorption wave meter is more sensitive and selective, but is still of limited use. Accurate field strength measurements can be made under laboratory conditions using a calibrated antenna and a calibrated measuring receiver.

A two-tone signal generator is used to inject a known signal into an SSB transmitter so that its output power can be measured. The PEP is twice the indicated output power from the transmitter.

Revision Questions

1 To extend the current range of a meter movement, a factor which must be known beforehand is the:

a. Full scale deflection voltage and coil internal resistance.
b. Maximum current-carrying capabilities of the meter movement.
c. Insulation resistance of the meter coil.
d. Maximum voltage the coil will take across its terminals.

2 To use the movement of a 0 to 50 µA meter to measure voltage in the range 0 to 10 kV, when the scale has been calibrated to read 0 to 100 V, use would be made of a:

a. Series resistor of approximately 200 MΩ
b. Series resistor of approximately 200 kΩ
c. Shunt resistor of approximately 200 MΩ
d. Shunt resistor of approximately 200 kΩ

3 One of the reasons for using a transistorised multimeter is its greater sensitivity. On a voltage scale, this means that:

a. It will load the circuit under test to a greater extent.
b. The circuit under test sees a much higher input impedance.
c. Greater sensivity allows the scale to be subdivided into smaller units.
d. The circuit under test will see a lower input impedance.

4 The basic instrument for measuring voltage and current is:

a. An oscilloscope.
b. A moving coil meter.
c. A field strength meter.
d. A tape measure.

5 What is a multimeter?

a. An instrument capable of reading voltage, current, and resistance.
b. An instrument capable of reading SWR and power.
c. An instrument capable of reading resistance, capacitance, and inductance.
d. An instrument capable of reading resistance and reactance.

6 How is a voltmeter typically connected to a circuit?

a. In series with the circuit.
b. In parallel with the circuit.
c. In quadrature with the circuit.
d. In phase with the circuit.

7 The range of an ammeter can be extended by adding resistance:

a. In series with the circuit under test.
b. In parallel with the circuit under test.
c. In series with the meter.
d. In parallel with the meter.

8 What is a dummy load?
a. An isotropic radiator.
b. A non-radiating load for a transmitter.
c. An antenna used as a reference for gain measurements.
d. The image of an antenna, located below ground.

9 What material may a dummy load, suitable for RF, be made of?
a. A wire-wound resistor.
b. A non-inductive resistor.
c. A diode and resistor combination.
d. A coil and capacitor combination.

10 What station accessory is used in place of an antenna during transmitter tests when no signal radiation is desired?
a. A Transmatch.
b. A dummy load.
c. A lowpass filter.
d. A decoupling resistor.

11 What is the purpose of a dummy load?
a. To allow off-the-air transmitter testing and adjustment.
b. To reduce output power for QRP operation.
c. To give comparative signal reports.
d. To allow Transmatch tuning without causing interference.

12 What is a marker generator?
a. A high-stability oscillator that generates a signal or series of signals from a single low-frequency signal source.
b. A low-stability oscillator that “sweeps” through a band of frequencies.
c. An oscillator often used in an aircraft to determine the craft's location relative to the inner and outer markers at airports.
d. A low-stability oscillator used for signal reception.

13 A dip oscillator is a type of:
a. RF signal generator.
b. Cathode ray oscilloscope.
c. Reflectometer.
d. RF wattmeter.

14 Which piece of test equipment contains horizontal and vertical channel amplifiers?
a. The ohmmeter.
b. The signal generator.
c. The ammeter.
d. The oscilloscope.

15 What is the best instrument for checking transmitted signal quality from a CW/SSB transmitter?
a. A monitor oscilloscope.
b. A field strength meter.
c. A sidetone monitor.
d. A diode probe and an audio amplifier.

16 When connecting a CRT oscilloscope to view the wave envelope pattern of an amplitude modulated transmitter, the following coupling method would be used:

a. Direct coupling.
b. Inductive coupling.
c. Driver input coupling.
d. Inductive coupling to the final tuned circuit.

17 The vertical deflection plates in a CRT oscilloscope may be used to measure the amplitude of a signal. This signal may be displayed in terms of:

a. Current.
b. Voltage.
c. Frequency.
d. Time.

18 What kind of input signal is used to test the Peak Envelope Power of an SSB transmitter while viewing the output on an oscilloscope?

a. Normal speech.
b. An audio frequency sine signal.
c. Two audio frequency sine signals.
d. An audio frequency square wave.

19 What can be determined by making a "two tone test" using an oscilloscope?

a. The percentage of frequency modulation.
b. The percentage of carrier phase shift.
c. The frequency deviation.
d. The amplifier PEP power output.

20 What is a reflectometer used for?

a. Checking the standing wave ratio.
b. Peaking a receiver’s sensitivity.
c. Transmitter noise figure measurements.
d. Measuring sunlight intensity.

Chapter 30: Digital Systems

30.1 Advantages of Digital Systems

As has become apparent in preceding chapters, analogue electronics can require lots of care to design, build and operate. Capacitors, inductors, resistors and semiconductors all have characteristics that are slightly variable and also change with time. As a result, the trend is increasingly to substitute analogue circuits with digital ones. Digital circuits have numerous advantages:

- They can be made much smaller and therefore much cheaper.
- They require no sensitive adjustment, either during manufacture or during operation.
- They are more robust against abuse, such as impact or vibration.
- Much of their operation can be automated.
- Design changes can be incorporated by software downloads rather than physical modification. In this way, apparatus can remain up to date for decades, even with the advent of new modulation techniques, regulations and procedures.

30.2 Principles of Digital Signal Processing

Signal processing is the action of modifying or enhancing one or more parameters of a signal to improve and select a wanted characteristic. In radio engineering this processing may entail any of the familiar operations such as amplification, modulation, filtering, mixing and detection.

All these functions may also be accomplished digitally with the aid of a computer. A simple example is a mixer. When discussing mixers, we stated that mixing is actually multiplication. This multiplication of two signals can easily be accomplished in a computer, yielding the same result. Amplification is likewise simple: Just multiply the signal by a bigger number, and you have a louder signal to work with.

Modern desktop and portable computers can operate at astonishing speeds. It is now possible to do all these calculations in a computer. However, specialised Digital Signal Processing (DSP) circuits can do the calculations much faster by using tricks like dedicated array processors, that can do thousands of multiplications or other operations simultaneously. Such DSP chips now make up much of the functioning of a typical modern amateur transceiver.

Although common microcomputers may be used for this purpose, specialised digital signal processing micros have been developed that execute the required operations much faster.

What is “Digital”?

The word “digital” is ubiquitous in our modern world, but few seem to understand exactly what it means. “Digit” is from the Latin word for “finger”. However, in this context it means a single number, such as 0, 1 or 2. “Digital” means “based on numbers”. The opposite is analogue, meaning something based on a continuously-variable physical quantity, such as voltage or current. No matter how small the difference between two voltages, there is always another voltage in between. Between 1 nV and 2 nV, there is 1½ nV. Between 1 nV and 1½ nV there is 1¼ nV. We could carry on forever; there is always another voltage in between.

Here’s another practical example: Day and night. Any specific moment can be described as “day” or “night”. If we only have these two terms to describe the time, we have to make a choice. It is either “day” or “night”. At noon and midnight, there is little fear of making an error. However, around sunrise and sunset there are periods when the distinction is not so

clear. Picking either description involves some error, as there is some “day-ness” and some “night-ness” at that time of day.

In digital systems, we work with discrete numbers, generally integers. Between two adjacent integers there is no other integer. Between 1 and 2 there is nothing. Sorry.

In computers (DSP chips are special-purpose computers!), we generally work with binary numbers. Just like normal folks work in decimal (base 10), computers work in binary (base 2). The decimal system needs 10 symbols to work (count them: *0 1 2 3 4 5 6 7 8* and *9*). The binary system needs only two (count them: *0* and *1*). The binary system is convenient because it is easy to represent two symbols in an electric circuit. You can decide that *off* (a voltage of 0 V) represents *0*, and *on* (a voltage of 5 V) is *1*. Accuracy is not all that important. In fact, you could decide that anything below 2½ V is going to be a *0*, and anything above 2½ V is a *1*. There is now lots of leeway for noise and error in the circuits, without introducing errors.

Using binary numbers, we can represent any integer quite easily. 1010_2 (*one zero one zero base two* or *one zero one zero binary*) is the same as 10_{10} (*ten base 10* or *10 decimal*). What is 101010_2? The answer, of course, is $42_{10.}$ To see how this representation works, we need to look at number systems.

Number Systems

The decimal system works with a *base* (or *radix*) of 10. A number such as 1234,567 only has meaning because we know what the base is. The meaning of every digit is determined by the base and by its position relative to the decimal separator (the comma). In this case, the number is represented by:

$$1 \times 1000 + 2 \times 100 + 3 \times 10 + 4 \times 1 + 5 \times 0{,}1 + 6 \times 0{,}01 + 7 \times 0{,}001$$

which can also be written as

$$1 \times 10^3 + 2 \times 10^2 + 3 \times 10^1 + 4 \times 10^0 + 5 \times 10^{-1} + 6 \times 10^{-2} + 7 \times 10^{-3}$$

If we don’t know the base, we also cannot make sense of the number.

In the binary system, we only have two symbols. For convenience, we’ll just use 0 and 1, using their traditional meanings, even though we are working in a different base. A binary number may consist of many digits. A binary digit is called a *bit*. Each bit is a *0* or a *1*. We indicate that we are working in binary (base 2) by inserting the base in subscript after the number.

Let’s try to interpret the following number:

$$101010{,}0110_2$$

The number can be expanded as:

$$1 \times 2^5 + 0 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 0 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} + 0 \times 2^{-4}$$

which we can rewrite, omitting the 0 terms, as

$$32 + 8 + 2 + 1/4 + 1/8 = 42{,}375_{10}$$

We retain the convention of showing the most significant digits on the left and the least significant digits on the right. In most computer systems, especially simple systems, binary

numbers are regarded as integers (i.e. nothing less than 1, or no digits after the comma). We refer to the most significant bit (MSB) on the left and the least significant bit (LSB) on the right. A group of four bits is called a *nibble* and a group of eight bits is called a *byte*. Several bytes together are known as a word. In modern desktop machines, 32-bit and 64-bit words are mostly used.

Although computers love binary numbers, the long strings of 0 and 1 symbols are hard for humans to read and interpret. To overcome this problem, programmers and engineers often use base 16. As $16 = 2^4$, binary numbers can be neatly broken up into nibbles and then converted into *hexadecimal* (base 16). Likewise, each hexadecimal digit can be converted to a nibble, allowing easy conversion to binary. To make up the sixteen symbols, we use 0 to 9 in their traditional meaning, then continue with A, B, C, D, E and F:

Binary nibble	Decimal number	Hexadecimal digit
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	10	A
1011	11	B
1100	12	C
1101	13	D
1110	14	E
1111	15	F

When using the hexadecimal notation, we denote the fact that it is hexadecimal with a subscript "16", such as $3FA4_{16}$. To make the notation easier, some programmers append an "H" to the number, as in 3FA4H. C programmers also use the notation 0x3FA4.

Because decimal numbers regularly need to be entered into computers, Binary Coded Decimal (BCD) was devised. Each decimal digit is simply encoded in binary and used as a nibble in the computer. Obviously, the resulting word is not a proper binary number and special handling was required to execute arithmetic on BCD words.

Logic Operations

Boolean logic was defined by George Boole (1815-1864). Just as the integers have operations (addition, subtraction, multiplication, inversion, division etc.), bits can also be inverted, added and multiplied. We cannot use the same definitions as for decimal, though, because the answer must always be a bit.

The rules of arithmetic in binary are:

A + 0 = A
A + 1 = 1
A ● 0 = 0
1 ● 1 = 1
-1 = 0
-0 = 1

A binary addition is said as *or*. A binary multiplication is said as *and*. A binary minus is said as *not*. To make sense of the rules of logic, perhaps we should work with "false" and "true" rather than with *0* and *1*.

Now, let's try to read the rules of binary arithmetic, given above, like logic statements:

True or false is true. False or false is false.
Anything or true is true.
Anything and false is false.
True and true is true.
Not-true is false.
Not-false is true.

Note that the definition of *or* is not quite aligned with our normal ideas. If I say "The sun is shining *or* it is raining", the answer is true if the sun is shining *or* it is raining, but not when both statements are true. In binary logic, an *or* is true even if both inputs are true.

Let's work through some examples:

"The moon is made of cheese" or "the moon is made of sand" is true.
"The moon is made of cheese" and "the moon is made of sand" is false.
Not-"the moon is made of cheese" is true.
Not-"the moon is made of sand" is false.

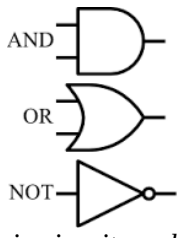
The behaviour of an operation can also be described by a *truth table*. Here is the truth table for *and*.

A	B	A ● B
0	0	0
0	1	0
1	0	0
1	1	1

*Truth table for **and***

The table basically says what we already know: *If A is true and B is true, A and B is true. Otherwise, if A and B are not both true, A and B is false*.

These simple operations are all that a computer has to work with. To facilitate the drawing of diagrams to demonstrate logic operations, we need some symbols:



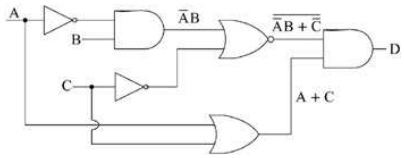
Logic circuit symbols

The top symbol is the ***and*** operation. If both inputs are high, the output is high. Otherwise, the output is low.

The second symbol is the ***or*** operation. If both inputs are low, the output is low. If either input (or both) is high, the output is high.

The bottom symbol is a ***not***. If the input is high, the output is low, and *vice versa*.

The diagram below shows a three-input circuit, with the output of each multi-input gate shown.



A simple logic diagram

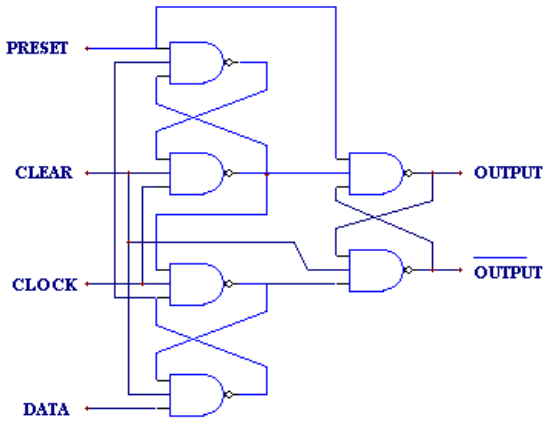
In the top line, the first gate provides not-A. The second gate provides not-A and B. The third gate has a small circle on the output, which represents an inverter. It is therefore a combination of an *or* and an *inverter*, called a *nor*. Its output is not (not-A and B or not-C). The final output of this circuit is D = not (not-A and B or not-C) and (A or C).

If you can read this diagram, you know enough to conquer some more complex examples that you will find in the real world. This kind of logic is known as *combination logic*. It simply takes all the inputs and calculates an output. If any input changes, the output may change. If all inputs remain constant, the output remains constant.

Incidentally, there are also gates which invert the output of an *and* gate, known as a *nand*. There is something that corresponds to our normal notion of or, called *exclusive or* or *eor*. The *eor* is false if both inputs are true (like our example of rain and clouds). One could simply say that *eor* is only true if only one input is true. There are gates that invert both inputs or invert only one input. There are gates that take more than one input, following the same rules we have described before.

There is also a second type of logic, known as *sequential logic*. This kind of logic takes into account things that happened in the past, such as the previous value of a specific input or output, and is often controlled by a clock. The clock is simply a square signal. The action normally happens when the clock changes, either going up or down. At this point, the new output is calculated based on the inputs and the previous values mentioned.

Sequential logic requires some memory. The basic building blocks of memory are called latches. One useful latch is called a D flip-flop. An example of a D flip-flop, built from nand gates, is shown below.



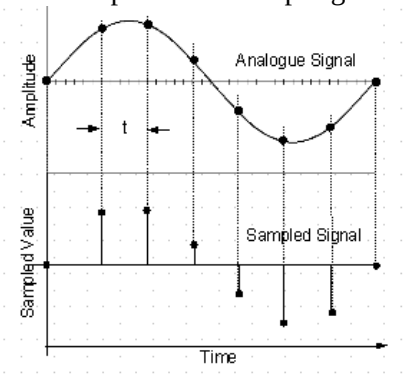
D flip-flop made from nand gates

At switch-on, the state of **Output** is undetermined so 1 is applied to the **Clear** input. **Output** becomes 0. The rising edge of the **Clock** transfers the value of **Data** to **Output**. Note that the two versions of **Output** are available; one is the inverted version of the other. If **Data** now changes state, **Output** will remain the same, at least until the next rising flank comes on **Clock**.

By connecting such, or other more complex flip-flops together and gating their outputs, complex logic sub-systems may be constructed. Such systems include counters, shift registers, arithmetic logic units, memories and, ultimately, complete computers.

Sampling

Before any signal processing can take place, the analogue signal has to be converted to a digital signal, which simply means it has to be converted into a bunch of numbers. This conversion is done by taking periodic samples of the analogue signal and storing the instantaneous values as numbers. The process of sampling is illustrated in the figure below.

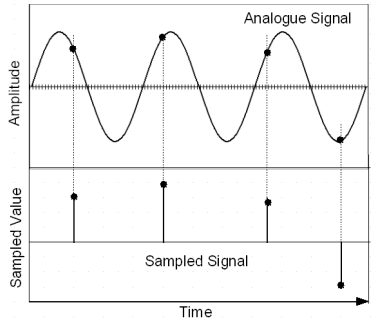


A sampling process

Note that the sampling period "t" is much shorter than the period of the sampled wave, about one-seventh in this particular example. Connecting the upper points of the samples produces a fair representation of the original signal at the same frequency.

The sampling rate is governed by the *Nyquist Sampling Theorem.* This theorem requires that at least two samples be taken during each cycle in order to faithfully reproduce the original signal.

What happens if we take less than two samples during a single cycle?
The reconstructed signal is still a sine signal, but of a different and lower frequency and does not represent the original signal. It is known as an *alias frequency*. In some applications deliberate use is made of this lower frequency and it is then called *undersampling.*



Effect of undersampling

In the case of a complex signal, it means that the sampling rate must be more than twice the highest frequency component contained in the signal.

In order to avoid aliasing, a lowpass or bandpass filter is inserted ahead of the sampling device. A practical example of this architecture is the sound card in a computer. To be able to adequately cover the audio frequency range of 20 Hz to 20 kHz, the sampling frequency is 42 kHz.

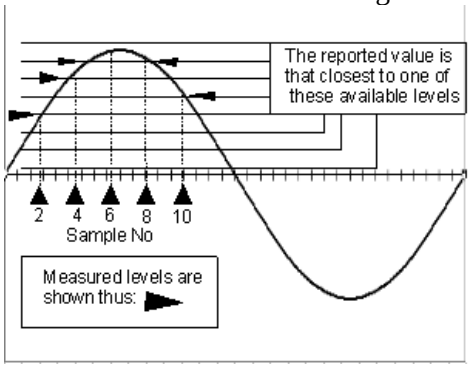
Analogue to Digital Conversion
The device used to convert the analogue signal to a sampled digital version is an a*nalogue-to-digital converter* (ADC). For each sample, the ADC produces a number that is directly proportional to the amplitude of the input voltage. The number of bits available in the binary word limits the number of discrete voltage levels that can be resolved. An 8-bit ADC can only resolve 256 levels (2^8) while a 12-bit unit can resolve 4096 (2^{12}) levels. The number of bits limits the resolution of the ADC as it can only report the analogue value to the nearest discrete level.

The difference between the actual and reported value is called the *quantisation error* and for a good ADC is ±½ the value of the Least Significant Bit (LSB) for that converter. For a 12-bit converter, the error is 1/8192, or about 0,01%. This error is pseudo-random and appears in each sample, depending on how close the sampling step was to the actual value measured. Once the numbers are converted back into a signal, the errors appear as *quantisation noise*.

Once the signal has been digitised, a digital process can be implemented to perform various functions amongst which are modulators, mixers, AGC systems and filters.

Try to imagine how an AM modulator and an AGC might work. Filters are not so simple, just yet.

The conversion process in an ADC is illustrated in the diagram below:



Rounding off during sampling

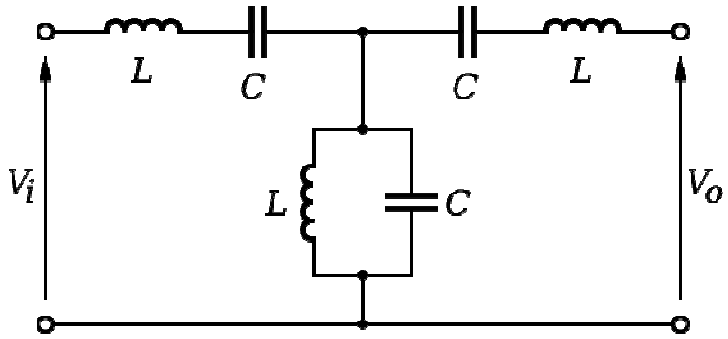
A further source of noise, especially in VHF signals, is due to *aperture jitter* which is caused by small variations in the sampling clock intervals. It is, however, much smaller than the quantisation noise. Yet another source of error is caused by the non-linearity of the conversion and this is typically ±1 or 2 bits over the entire range.

Digital to Analogue Conversion

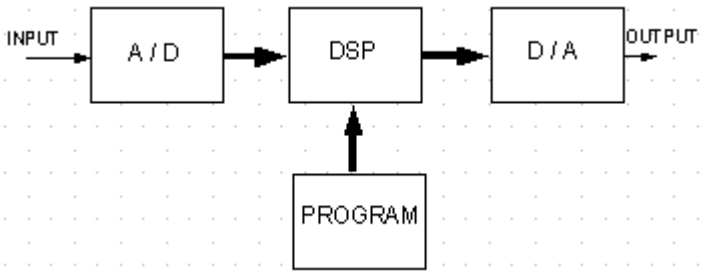
After the digital processing has been completed, the resulting string of numbers must be converted back to an analogue value. A *digital-to-analogue converter* (DAC) does this job. The device accepts a series of numbers as input and, at the application of a clock signal, produces the corresponding analogue value. The discrete nature of the digital input results in a stepped analogue output that may be smoothed by filtering.

30.3 Digital Filters

As mentioned previously, DSP can be used to implement a number of analogue functions. The figures below show the alternate implementations of a bandpass filter.



An analogue bandpass filter



A digital bandpass filter

Digital filters are implemented in one of two ways, called IIR and FIR filters. The difference between them is that IIR (*Infinite Impulse Response*) filters involve results of previous calculations (feedback) while FIR (*Finite Impulse Response*) filters do not. The name IIR refers to the response of the filter to a single unity amplitude sample pulse. The name is somewhat of a misnomer as it would indicate that the output of the filter will last forever while it actually gets smaller until it falls below the resolution of the processor.

Using fancy mathematics called the Fourier Transform, the pulse response of a filter can be converted to its frequency response. Just like we can draw a graph showing the frequency response of an analogue filter, we can draw a graph of the frequency response of an IIR or FIR filter.

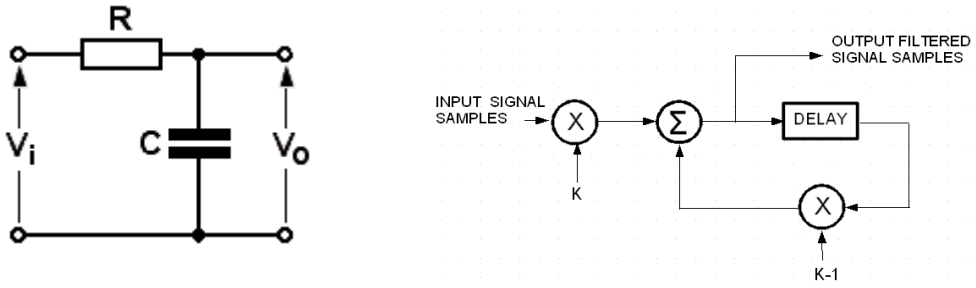
In the discussions that follow, we will use the following symbols in order to explain the operation of digital filters.

Multiplication of Inputs Summation of Inputs One Sample Period Delay

Symbols for DSP operations

IIR Filters

The simplest IIR filter is the implementation of the RC lowpass filter shown below. The block diagram shows the simple IIR filter that has the same response as the R-C filter.



Analogue and IIR versions of a simple RC lowpass filter

If we call the digital input sample x_i and the filter output y_i, our filter consists of the single calculation:

$$y_i = K\,x_i + (1 - K)\,y_{i-1}$$

where K is a constant between 0 and 1 but typically 0,001 or less.

This equation simply means:

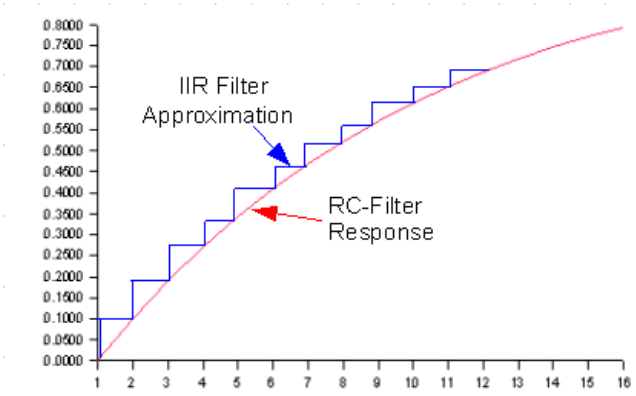
Output number i is equal to input number i plus slightly less than the previous output, number i – 1.

Demonstrating how it works

Using the above equation we can now calculate the operation of this filter for the first few terms as the input rises from *0* to *1*. Assume that the output is 0 when we start and choose K = 0,1 to make things happen faster.

New Input, x_i	K x_i	(1-K) y_{i-1}	New Output
0,0	0,0	0,0	0,0
1,0	0,1	0,0	0,1
1,0	0,1	0,09	0,19
1,0	0,1	0,171	0,271
1,0	0,1	0,244	0,344

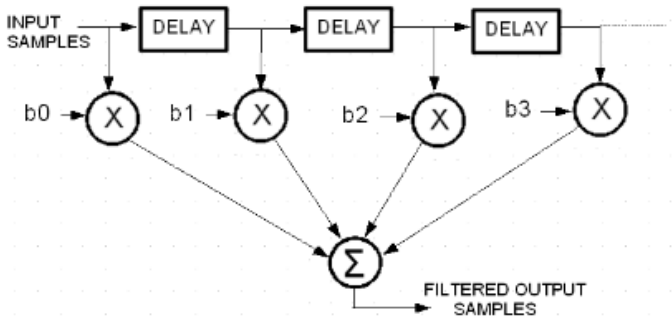
The calculation shows that the output is growing towards 1 but with a smaller step after each new input. This output is the same exponential growth as that of an R-C circuit. The figure below shows the charging characteristics of the RC filter as well as that of the digital circuit.



FIR Filters

For more complex filters it is often desirable to use the FIR filter, standing for *Finite Impulse Response*. These filters don't use the previous outputs of the filter computation, instead using the current input along with many of the previous inputs.

DSP implementation of the FIR filter is very simple as shown in the block diagram below:



FIR Filter

The input signal is available in digital format from the A/D converter. A delay line consists of locations in memory where previous samples are stored. Each time a new sample arrives, it is placed in the beginning of the delay-line memory. Multiplying all the samples by constant numbers and then adding them together forms new outputs. The constant multiplier numbers (b1, b2, b3, etc.) are referred to as the *FIR coefficients*, or *tap weights*. The filter design consists of choosing these coefficients to suit the particular application. As

with analogue filters, there are tradeoffs between the number of coefficients, passband ripple and the out-of-band rejection.

The FIR structure can be used to form filters that are highly selective to the frequency of a sinusoidal input signal. All of the response characteristics of LC filters, such as Butterworth, Chebyshev and others are possible with the FIR filter.

30.4 Direct Digital Synthesis

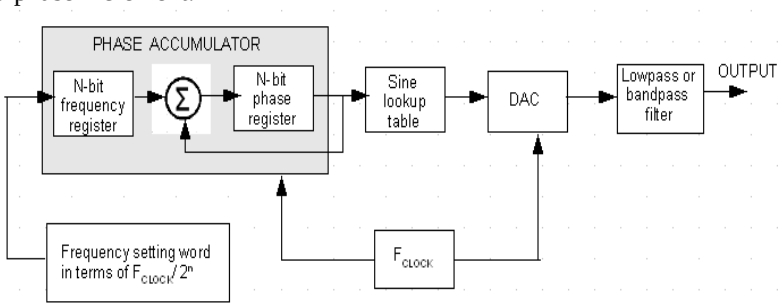
Direct Digital Synthesis (DDS) is the process of generating arbitrary waveforms by means of digital methods. The word "direct" refers to the fact that no feedback is used in the basic process. It is based on the fact that, for a sine signal, there is a fixed relationship between the phase and the amplitude values of the signal.

In a previous discussion, we dealt with the sine function and how it is derived. We looked at the vertical component of a rotating wheel, through a rotation from 0° to 360°. We also commented that the application of the sine function in electric circuits normally describes the function $V = VPeak \sin 2\pi ft$.

The values of sine are well known and can be listed in a lookup table. If we want to produce a sine function, we can simply read the values of successive samples from the table and put them into the ADC one by one.

Creating a sine signal by Direct Digital Synthesis (DDS)

The diagram shows a DDS system to generate a sine signal at a given frequency by digital integration of a phase increment.



Block diagram of a DDS system

The controller determines which frequency needs to be produced, based on user input and system architecture. It then determines an increment to be added to the phase register in each cycle. When the clock is cycled, the increment is added to the phase register, producing a new phase angle. Remember that the phase returns to 0 if 360° is reached, so the register rolls over when it reaches this value. Once the phase register contains the correct phase value for the new sample, the output value for the sine function is looked up. This value is fed to the DAC, which produces an output voltage corresponding to the desired momentary output value. Because this output value may contain steps due to quantisation error, it must be filtered before being used as input to mixers or other components.

If *N* bits are available to the phase accumulator, the output is given by:

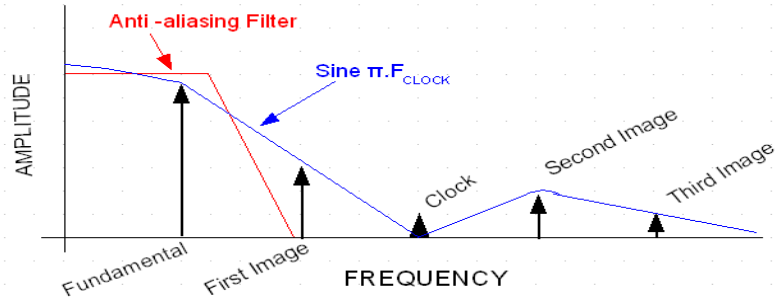
$$F_{OUT} = (\text{Frequency Setting Word}) \times F_{Clock} / 2^N$$

The lowest possible output frequency as well as the frequency increment is $F_{Clock} \div 2^N$ which occurs when the FSW is 1. Using a clock frequency of 100 MHz and $N = 32$ gives rise to a frequency increment of 0,0232 Hz!

As the clock frequency is usually derived from a stable crystal oscillator, the DDS can deliver output frequencies with a very high stability and resolution. The maximum output frequency is usually

limited to about half of the clock frequency and sometimes even lower. The stepwise output also generates unwanted signals of which most may be eliminated by the judicious choice of the clock frequency and output filters.

The diagram below shows the spectrum of the stepwise output before filtering:



DDS Output Spectrum

30.5 The Fourier Transform

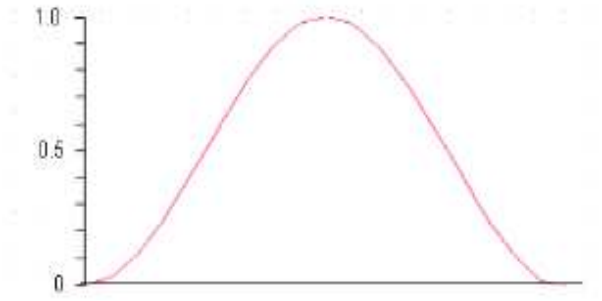
A Fourier transform is a mathematical technique for determining the frequency content of a signal. It was originally developed by Joseph Fourier (1768-1830) for continuous signals. For sampled signals, such as used in DSP, a variant of the transform called the *discrete Fourier transform* (DFT) is used. Given a a block of *N* input samples of a signal, the DFT will produce an output showing which sine function frequencies are included in that signal.

As the transform makes use of complex sinusoids, it is very demanding on computational processing. Using the direct method of computation requires *N* complex multiplications and additions per output frequency. For *N* such output frequencies, the computation effort is proportional to N^2 After realising that the complex sinusoid was periodic with *N*, mathematicians (notably Carl Runge 1856-1927) further simplified the method and reduced the computational effort. The machines of the day were too slow to implement the algorithm in real time, so the DFT lay dormant until it was revived by Cooley and Tukey during the 1960s. It is known as the *Fast Fourier Transform* (FFT) and is in common use in many consumer electronic devices.

Discontinuities and Windowing

When using a system of blocks of data for analysis, discontinuities exist at both the beginning and end of the blocks which cause unexpected spectral components to appear in the output. One way of alleviating this problem is to increase the number of discrete output frequencies, but this increase unfortunately complicates the decision as to which output frequency the result is to be allocated.

A technique called *windowing* addresses the problem by multiplying the data block by a *window function* which removes the sharp transitions in the envelope. The samples are then processed in the normal way. The best known and used window functions are the *Triangular, Blackman, Hamming, Hanning* and the *Rectangular*. Using a rectangular window is the same as not using a window at all while all the others have a shaping effect. The spectra of these windows all resemble that of a lowpass filter and are often used to design such filters. The diagram below shows the shape of the Hanning window.



Hanning (or cosine bell or cos²) window

30.6 Convolution

Convolution is an operation that calculates a signal from two other signals, comparing one signal to a time-inverted version of the other signal.

To determine the output of a network, the frequency spectrum of the input is multiplied by the frequency response of the network to obtain the frequency spectrum of the output.

The equivalent operation in the time domain is to convolve the input signal with the impulse response of the network to obtain the output signal.

Convolution is thus the time-domain equivalent of frequency-domain multiplication.

In some cases, convolution is the right choice to simplify calculations. In other cases, it is more efficient to do an FFT on the input signal, then to multiply with the frequency response of the circuit and then to do an inverse FFT to obtain the output signal.

30.7 SDR Platforms

Increasingly, the trend in radio manufacturing will be towards software-defined radios (SDR). A generic radio featuring minimal analogue components and a powerful DSP will be made to do whatever is needed by the specific user. Some VHF radio manufacturers are already making a single radio that can be programmed to use a specific modulation mode on specific channels with specific tone combinations for a specific user. Only one item is kept in stock, but it can be reconfigured for each user.

A popular experimental platform for SDR is Linrad. The software is available as a free download and runs under Linux, without any licence or copyright restrictions. Linrad can be used with an ADC directly connected to the antenna, or with a receiver and a sound card. It can also function as a transmitter, using a suitable DAC. The software does not require elaborate hardware, and can run on obsolete previous-generation computers. It provides great opportunities for experimentation at low cost.

Summary

Digital circuits have become ubiquitous in modern electronics. They are cheaper, smaller, easier to make and more stable than their analogue counterparts.

All the required operations in radios, including amplification, modulation, filtering, mixing and detection, can be done in digital systems using numbers rather than voltages.

Specialised Digital Signal Processing (DSP) circuits can do the calculations very fast by using tricks like dedicated array processors. Such DSP chips now make up much of the functioning of a typical modern amateur transceiver.

In binary systems, there are only two states (0 or 1). Binary numbers lend themselves well to computers, where these symbols are simply represented by *on* and *off*.

The most significant bit (MSB) is written on the left and the least significant bit (LSB) on the right. A group of four bits is called a *nibble* and a group of eight bits is called a *byte*. Several bytes together are known as a word. In modern desktop machines, 32-bit and 64-bit words are mostly used.

Hexadecimal notation is easier to read than binary, and can readily be converted to and from binary. We denote the fact that it is hexadecimal with a subscript "16", such as $3FA4_{16}$. To make the notation easier, some programmers append an "H" to the number, as in 3FA4H. C programmers also use the notation 0x3FA4.

In Binary Coded Decimal (BCD), each decimal digit is simply encoded in binary and used as a nibble in the computer. Obviously, the resulting word is not a proper binary number and special handling was required to execute arithmetic on BCD words.

In Boolean algebra, we use *or*, *and* and *not* to complete logic operations. These operations and combinations such as *nor*, *nand* and *eor* enable us to make complex *combination logic*.

In *sequential logic*, the past state of inputs and outputs may also influence the state of the output. Sequential logic uses memory elements such as D latches to retain past states.

An Analogue to Digital Converter (ADC) converts analogue signals to numbers. The Nyquist sampling theorem states that at least two samples per cycle must be taken if the signal is to be reproduced faithfully. *Quantisation noise* is introduced because the number does not necessarily represent the exact value of the voltage sampled at that moment.

DSP is done simply on the numbers inside the computer. The Digital to Analogue Converter (DAC) converts these numbers back into analogue signals. Because the output contains steps due to the discontinuous nature of digital signals, some filtering may be required after the DAC.

Most filters can be implemented in DSP systems just like in LC, RC or LR circuits. Digital filters can be implemented as Infinite Impulse Response (IIR) or Finite Impulse Response (FIR) filters. The IIR filter uses the current input and previous states to calculate the output, while the FIR filter uses the current input and a few recent versions of the input to calculate the output.

Direct Digital Synthesis (DDS) uses a lookup table of the values of the sine function to calculate the output voltage at any given moment. The current phase is calculated from the previous phase and an increment, which depends on the clock period and the frequency being synthesised. DDS provides very precise and accurate frequency synthesis, provided a suitably accurate timebase is used.

The Fourier Transform converts a signal into a frequency spectrum. For reasons of computational efficiency, most systems use a Fast Fourier Transform (FFT) algorithm to convert the time signal to a spectrum. Windowing algorithms are used to address bad effects of finite sample lengths.

Convolution is the time-domain equivalent of frequency-domain multiplication. Convolution sometimes simplifies calculations, yielding a more efficient answer than using FFT and reverse FFT to obtain the same result.

The future of radio is in SDR, using generic computers to implement entire radios with only simple ADC and DAC circuits. Linrad offers a free platform for experimenting with SDR.

Revision Questions

1 Digital circuits have become popular in consumer electronics and amateur radio because they are:

a. Easier to manufacture and adjust than analogue circuits.
b. Smaller than analogue RF circuits.
c. Cheaper than analogue RF circuits.
d. All of the above.

2 Digital systems are different from analogue systems because they:

a. Use digital displays instead of mechanical meters.
b. Work with keyboards instead of microphones.
c. Display received messages on a screen rather than playing them through a speaker.
d. Use computers to implement traditional circuitry such as mixers, amplifiers and oscillators.

3 A word can represent:

a. 16 bits.
b. 32 bits.
c. 64 bits.
d. Any multiple of eight bits.

4 Both hexadecimal numbers and BCD are used in computer systems. Both use nibbles to encode their content. The difference between BCD and hexadecimal is that:

a. BCD is harder to convert to decimal numbers.
b. Hexadecimal is harder to convert to binary numbers.
c. Hexadecimal is easier to convert to binary numbers.
d. BCD takes up less space.

5 Boolean algebra is different from classic algebra because:

a. It does not feature calculations like addition and multiplication.
b. It can only be used with two variables at a time.
c. It operates only on binary numbers.
d. It is old-fashioned and no longer in use.

6 The following is *not* a hexadecimal number:

a. 0x3A7B.
b. $3C8F_{16}$.
c. 3C8FH.
d. 0x3G2L.

7 Combination logic makes use of:

a. The inputs plus the state of the circuit at that moment.
b. The inputs plus state of the circuit at that moment plus a clock.
c. The state of all the inputs at that moment.
d. The state of all the inputs at that moment plus a clock.

8 Sequential logic makes use of:

a. The inputs plus the state of the circuit at that moment.
b. The inputs plus state of the circuit at that moment plus a clock.
c. The state of all the inputs at that moment.
d. The state of all the inputs at that moment plus a clock.

9 Nyquist stated that, to retain the characteristics of a signal well enough, analogue signals must be sampled:

a. With at least two voltage levels.
b. At least twice per cycle.
c. With at least three voltage levels.
d. At least three times per cycle.

10 Quantisation noise is caused by:

a. Hot semiconductor components.
b. Certain regions in the galaxy.
c. The fact that the sampling level does not correspond exactly with the voltage being sampled.
d. The fact that the sample is not taken at exactly the right moment.

11 The difference between IIR filters and FIR filters is that:

a. IIR is digital and FIR is analogue.
b. FIR is digital and IIR is analogue.
c. IIR uses current and past inputs while FIR uses the input and past outputs.
d. FIR uses current and past inputs while IIR uses the input and past outputs..

12 To produce a sine signal, DDS uses:

a. A combination of gates and flip-flops.
b. A lookup table and a clock.
c. Many different gates and inverters.
d. Many latches, including D flip-flops.

13 The FFT is used to:

a. Measure the frequency content of a signal.
b. Calculate the different frequency components of a signal efficiently.
c. Measure the frequency response of a circuit.
d. Calculate the frequency response of a circuit efficiently.

14 Convolution is used to:

a. Multiply two signals together.
b. Multiply two frequency spectra together.
c. Avoid having to calculate a bunch of FFTs and inverse FFTs.
d. Avoid having to calculate a bunch of MSBs and LSBs.

15 SDR is:

a. A way to transfer software via radio.
b. A radio that uses a computer as a user interface.
c. A radio that is completely implemented in software, with simple ADC and DAC circuits.
d. A Sophisticated Digital Radar.

16 A flexible platform for experimenting with SDR is:

a. Linux.
b. Ubuntu.
c. Linrad.
d. WSJT.

Chapter 31: Digital Communication Modes

31.1 Practical Implementation of Digital Communications

In addition to the normal voice communication modes, many different digital modes are also in use in amateur radio. These modes can broadly be divided into two categories: *text modes* and *image modes.* Most of these modes can now be implemented in software, using the sound card in a PC and a normal speech communication medium, such as an SSB speech channel.

The input to the sound card is coupled to the audio output of the transceiver via the loudspeaker output or any other dedicated audio output connector. The speaker output is normally controlled by the volume knob, so most radios provide a constant-level output especially for sound cards. Audio from the sound card output is fed to the microphone or other modulation input of the radio. Provision is also made to key the transmitter from one of the computer's ports. Algorithms within the software control the signal processor in the sound card to process the audio frequencies and phases of the incoming and outgoing signals.

31.2 Digital Modulation

Digital modulation is not dissimilar from analogue modulation methods as already discussed. One can simply think of a simpler waveform entering a microphone and being modulated in exactly the same way as speech.

Modulating a square signal onto an AM carrier produces CW. Modulating a square signal onto FM produces FSK.

Modern digital communications systems often use this simple approach. A generic computer with a generic sound card is connected to a standard SSB transmitter. Using suitable software, digital signals are transmitted. At the receiving end, a standard SSB receiver is used to recover the signals, and a standard sound card demodulates the signals and displays them on the screen, or renders them as speech, or interprets them and shows them as a waterfall or in some other form acceptable to the operator.

Most amateur radio contest logging software now integrates with digital-mode software, allowing the operator to log contacts directly without having to invoke any special software or hardware. This simplicity and economy has contributed greatly to the soaring popularity of digital modes. According to ClubLog statistics, 2015 is the first year in which more data contacts than phone contacts were made world wide! CW remains the mode in which the largest number of contacts is made (at least by people uploading to ClubLog).

31.3 Text Modes

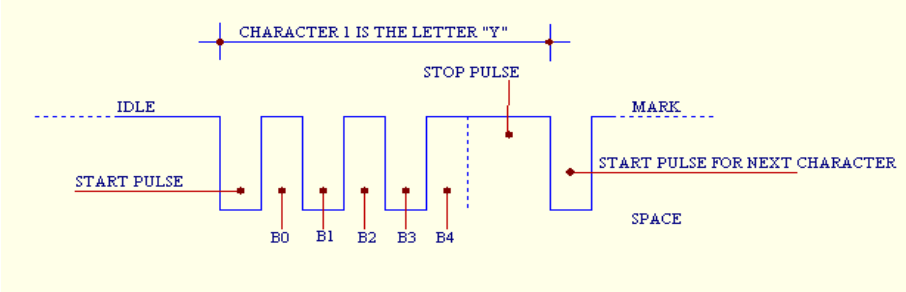
Morse Telegraphy

Morse telegraphy is the original mode for both amateur and commercial radio communication and operates by on-off keying of the transmitter carrier wave. It occupies a very narrow bandwidth and is readable even under very marginal conditions. The signals are human-readable up to about 60 words per minute.

Until recently, humans had the edge when reading Morse code. With the advent of CW Skimmer, all that has changed. CW Skimmer can read all the Morse code in an entire amateur band simultaneously, using a sound card and a PC.

When operating Morse with a computer, the CW transmitter must be keyed directly to take advantage of the transceiver's keying waveform shaping and sophisticated filtering. Some amateurs attempt to feed Morse tones into an SSB transmitter input. This practice is illegal,

Due to the mechanical nature of the early machines the signalling rates are not very high and vary between 65 and 133 words per minute. It is common to measure the signaling rate of RTTY in *"baud"*. The figure below shows the formation of the letter "Y ". Typically, the pulse length is 22 ms which equates to 45,45 Bd, a common signaling rate. In European (and South African) landlines a signaling rate of 50 Bd was common.



Character "Y" in Baudot (ITA2)

as SSB is not allowed in the Morse subbands. It is also bad practice, as the technique often results in hum, clicks and pops being transmitted on the air.

Although Morse proficiency is no longer required to get a licence, CW is more popular than ever. Major CW contests attract tens of thousands of participants.

Radio Teletype (RTTY)

RTTY is one of the first data communication modes that came into widespread use. FSK with a shift of 170 Hz is used, with a data rate of 45,45 Bd. RTTY is widely used to casual keyboard-to-keyboard chatting and for working DX. Because it is slow and error prone, it is seldom used for data transmission.

The five-bit Baudot code could only encode $2^5 = 32$ symbols. There were not enough symbols to cover the 26 letters, the 10 figures and the punctuation marks. The problem was solved by using two of the symbols to select one of two modes: Letters mode and Figures mode. The same symbol could then be interpreted as either a letter or a figure, providing enough symbols to for letters, figures and punctuation. The Baudot code is also known as the *International Telegraph Alphabet No 2* (ITA2).

Early systems used *Frequency-shift Keying* (FSK), moving the carrier frequency by 170 or 850 Hz. Modern systems employ *Audio Frequency-shift Keying* (AFSK) using an audio tone of 2125 Hz for a *mark* level and 2295 Hz for the *space* level (it is a binary system), to retain the same shift. It is used mostly on HF, but also occasionally on VHF.

AMTOR

Amtor is a development of RTTY. It is no longer in use, having been displaced by more sophisticated protocols and by RTTY itself.

Amtor is based on a system devised in the Maritime Mobile Service to improve communications between stations using RTTY. The system overcomes some of the problems experienced by RTTY when signal fading and noise occurred. The system converts the 5-bit code to a 7-bit code in such a manner that there are four mark and three space bits in every character. There are two modes that are commonly used in AMTOR. *Mode A* uses automatic repeat request in which the receiving station acknowledges the received characters by checking for the correct 4/3 bit ratio or else calls for a repeat. *Mode B* uses a simple forward error correction by sending each character twice.

AMTOR is a relic of the past, superceded by more sophisticated error correcting modes, and by its predecessor RTTY, that it intended to displace.

ASCII

The *American Standard Code for Information Interchange* is commonly used in computers, communication systems and related equipment.

ASCII uses a seven-bit code that can accommodate $2^7 = 128$ symbols. Although it is not strictly part of the ASCII standard, an eighth bit may be added as a parity (error checking) bit. By using other methods for error checking, the eighth bit may be used to extend the set of ASCII symbols to 256, making provision for some non-standard characters.

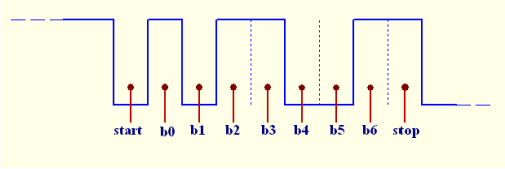
In radio communication, the ASCII character set is mostly used for serial data transmissions. These transmissions may be either synchronous or asynchronous. For synchronous transmission the character code is sent on its own while for asynchronous transmission a start bit and one or two stop bits are added. The standards for serial transmissions require that the character be transmitted with the least significant bit first and that the start and stop pulse duration be the same as that of the information pulse.

The signalling rates differ depending on the medium used but in amateur radio the following are most widely used: 75, 110, 150, 300, 600, 1200, 2400, 4800, 9600, 16 000, 19 200 and 56 000 b/s (bits per second). Signaling speed is often quoted in baud (Bd) which is equal to one discrete condition or event per second.

Some digital modulation methods have more than the normal two states. In the dibit modulation method, two ASCII bits are sampled at a time and have values of *00*, *01*, *10* and *11*. The four-phase modulation method assigns phase of 0º, 90º, 180º and 270º respectively. For this type of phase modulation the signaling speed in baud is half the transfer rate in b/s. Many modern modulation techniques have been developed that exploit this multi-bit encoding.

ASCII code example.

As an example the code for "M" is *1001101*. A pulse sequence showing the serial transmission of the letter "M" using ASCII is shown at right.

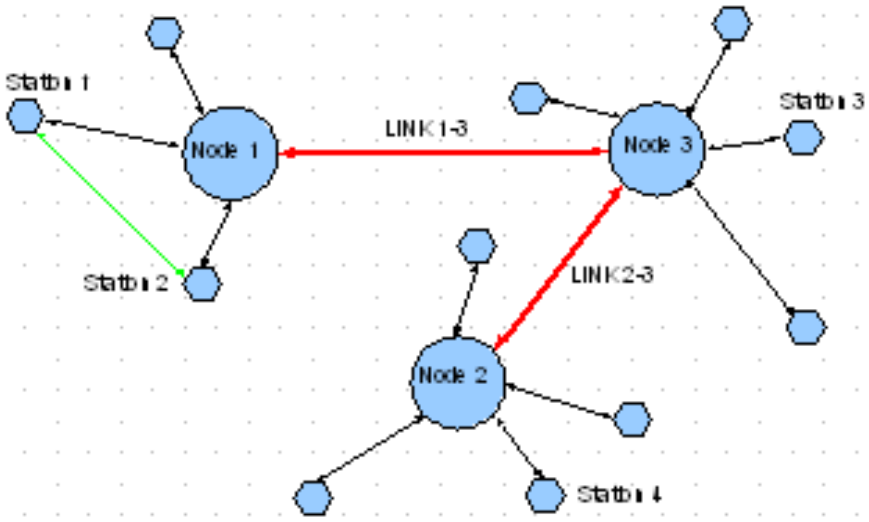


ASCII character "M", 7-bit, 1 stop, no parity

Packet Radio

Data communications is telecommunications between computers. *Packet switching* is a form of data communications that transfers data by subdividing it into "packets", and *packet radio* is packet switching using radio. (Steve Ford, WB8IMY)

From this definition, it follows that amateur packet radio is the communication between computers using amateur radio stations and that the computer operators are radio amateurs. The system uses only a single channel to service multiple communications simultaneously and is readily connected into networks that cover a given area. At the hub (or node) of each area we find a packet bulletin-board system (PBBS) which can store data and also forward it to other hubs. A possible configuration is shown below.



Layout of a Packet Network

Stations may communicate directly or via one or more nodes. Nodes can be connected in national as well as international network configurations. In order to maintain compatibility and also ensure virtually error free data communications, packet radio uses a modified CCITT (International Telephone and Telegraph Consultative Committee) recommendation X-25 protocol called AX-25. The main difference is that the address frame in AX-25 can accommodate amateur callsigns and has an added unnumbered information (UI) frame. This protocol specifies the format of a packet radio frame and the action a station must take when it transmits or receives such a frame.

FLAG	*ADDRESS*	*CONTROL*	*INFO/MSG*	*FCS*	*FLAG*
0111111	14 to 70 bytes	1 byte	Message or data up to 256 characters	Frame Check Sequence 2 bytes	0111111

Destination	Source	Digipeaters, Nodes, Paths
ZS6XXX-9	ZS5XYZ	

Format of an AX-25 frame

The heart of any packet radio system is the *Terminal Node Controller* or TNC. The function of the TNC is to take the asynchronous data from the computer or terminal (usually in ASCII form) and assemble it into packets or frames. These frames are then passed on to a modem for conversion into audio tones which in turn are fed to the radio transmitter. During reception the reverse takes place.

It is not always necessary to use a TNC as several computer programmes have been developed that use only a simple modem or sound card while the assembly and disassembly of the frames are done by the computer.

Most TNCs can also be used to re-transmit the received packets to other stations; in a process called *digipeating*. These stations do not add any information and merely re-transmit any frame that contains their callsign in the *digipeat* portion of the address field in the frame. In order to handle network operation, the native firmware in the TNC-2 is replaced with firmware called NET/ROM. This firmware supports the network and transport layers (levels 3 and 4) of the packet-radio network.

Packet bulletin boards, normally referred to as a BBS, have facilities to store messages which may be retrieved at a later stage by the addressee or, if placed in a public area, by any station requesting such a file. Examples of the latter are newsletters or other shared information.

One of the most popular applications of packet radio is TCP/IP which stands for *Transmission Control Protocol/Internet Protocol.* TCP/IP is the same protocol used to run the Internet, and actually a set of several protocols that provide a flexible and adaptable means of networking. The *Telnet* protocol allows for a chat session, while the *File Transfer Protocol (FTP)* allows the transfer of files between stations. A large number of software sets for TCP/IP is available based on the original NOSNET by Phil Karn, KA9Q.

APRS

Automatic Position Reporting System (APRS) provides real-time tracking of a vehicle or individual with a GPS receiver and an APRS radio.

The APRS or *Automatic Position Reporting System* was developed by Bob Bruninga, WB4APR, to overcome the limitations of packet radio when used for realtime communications. The limitations are mainly that a permanent link between all stations would be required for packet radio. APRS uses UI frames so that any number of stations may participate in the exchange of information.

For many events, the position of the radio is important and provision has been made to add a *Global Positioning System* (GPS) receiver to the system enabling moving targets to be tracked. This tracking is extremely useful for search and rescue operations as control stations are able to track the position of rescue workers and vehicles on a computer map in the control center. Some systems also allow the data from digital weather stations to be added to the transmitted data.

PSK31

PSK31 is a realtime keyboard-to-keyboard protocol, using PSK at 31,25 Bd. It can be used with weak signals and in noisy conditions.

PSK31 was developed by Peter Martinez, G3PLX, and derives its name from the fact that it uses phase shift keying at a baud rate of 31,25 baud. A new variable length code was developed for PSK in which the most used characters have shorter codes. The mode is very robust and can maintain communication even under very adverse propagation conditions.

Error correction was added to PSK by using *quadrature phase shift keying* (QPSK) and a convolutional encoder to generate one of four different phase shifts that represent five successive data bits. A Viterbi decoder is used at the receiving end to correct errors. This decoder tracks the 32 possible (5 bit) sequences, retaining only the most likely ones while discarding the other.

WSJT

Nobel physics prize laureate Joe Taylor K1JT has assembled a suite of different modulation modes in a single package called Weak Signal by Joe Taylor (WSJT). Many different modulation modes are included, all using variations of PSK and FSK plus multiple

redundancy and very slow baud rates to achieve spectacular weak-signal performance. Using WSJT—which can be downloaded for free—and a normal sound card, anyone can enjoy weak-signal communications under conditions so poor that the operator may not even be able to hear a signal. The different modes are optimised for different applications. JT65B, for example, is optimised for EME (earth-moon-earth) communications at 144 MHz, and has become the de facto standard for that mode. Using JT65B, EME is now within reach of almost anyone.

Detractors point out that their interest in ham radio is largely driven by what they hear. They wonder what the attraction is when your PC is contacting another PC, and the operator cannot even hear the weak signals…

FSK441 is a fast digital mode for meteor scatter. JT44 is a slow digital mode for troposcatter and earth-moon-earth (EME) communication. An EME echo mode is included for measuring your echoes from the moon. The JT65 series is optimised for EME. JT65B is specifically for the 2 m band, providing the majority of EME activity world wide.

CLOVER

Clover is an adaptive HF data system that adjusts itself to propagation conditions. Derivatives of the system are in use by maritime networks, the Civil Air Patrol and many amateurs. Bandwidth is 500 Hz and data rates around 250 Bd.

Clover was developed by Ray Petit, W7GHM in 1990. It uses a four-tone modulation scheme and allows different modulation formats to be selected manually or automatically depending on the prevailing signal conditions.

PACTOR

PACTOR uses a packet-based protocol with AMTOR-like encoding to attempt to combine the advantages of AMTOR and packet radio. It provides error checking to achieve very low error rates. It achieves data rates of 100 to 200 Bd in a normal HF bandwidth of 500 Hz.

Some benefits of PACTOR include: Operation at 100 or 200 baud, depending on path conditions; a 16-it cyclic redundancy check to provide near error-free operation; memory ARQ in which the controller is able to combine parts of successive blocks to eliminate errors; and selective use of IRA data compression.

Originally copyrighted by SCS GmbH, the mode was released into the public domain in 1991. Subsequent versions –II and –III with increased speed and other capabilities remain proprietary.

You can listen to samples of various digital modes including Pactor©-I, -II and –III at http://www.wb8nut.com/digital.html.

31.4 Image Modes

Facsimile

This is the name for methods that are used to transmit very high resolution still pictures using voice bandwidth radio channels. It is the oldest of the image transmitting methods and has been the main method used to transmit weather charts and newspaper photos by radio. It is also used by polar orbiting weather satellites to transmit their ground and cloud images to earth. FAX transmissions are made up of 800 to 1600 scanning lines, which provide higher resolution than analogue TV.

Modern amateur radio applications are based on software, using the sound card in a PC to decode the facsimile transmission and displays the image on the computer monitor. The high resolution is achieved by slowing down the data rate resulting in transmission times of four to ten minutes per image.

Slow-Scan Television (SSTV)

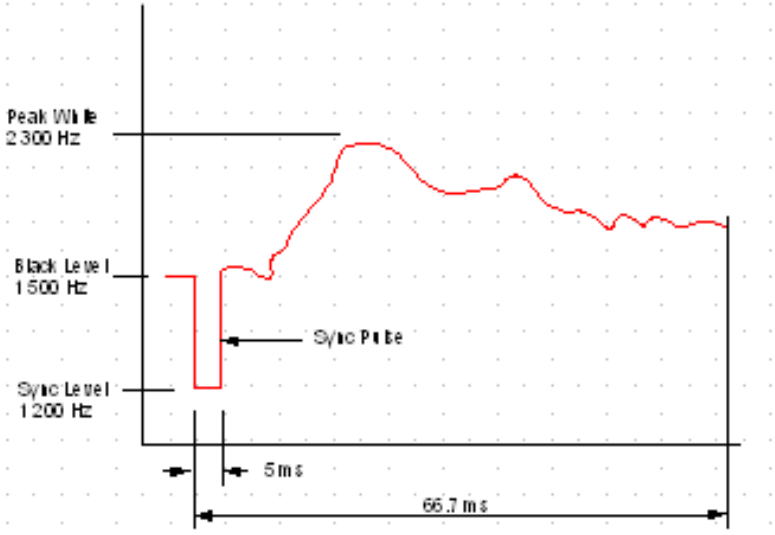
As the name implies, this mode differs from the normal commercial television with respect to the scanning rate of the picture frames. Home TV in South Africa (PAL) produces 25 complete images per second, with 625 lines each, and occupies a bandwidth in excess of 6 MHz. SSTV may take several minutes to complete a single image, but fits into normal speech channels, less than 3 kHz wide. The simplest monochrome modes complete a picture in about 8 s.

In order to transmit an image, very accurate and low distortion audio tones in the range of 1500 to 2300 Hz are used to represent the picture element (pixel) intensities. A 1200 Hz tone is used for synchronisation. These audio tones are then transmitted through a standard SSB speech channel. On the receiving side, the audio tones are decoded and turned into a picture. Most modern systems make use of the computer sound card and appropriate software to encode and decode the picture.

More than 40 different SSTV modes exist. The most popular modes are Scottie S1 and Martin M1, with various modes in the Scottie, Martin, AVT and Robot families being popular in various parts of the world.

Generally, a 5 ms synchronisation pulse indicates the start of a line and a longer pulse at the same frequency the start of a new frame, which is made up of 120 lines.

The same standard was transferred to colour images by placing red, green and blue filters in front of the camera and sending each colour image separately and assembling them at the receiving end. It was known as the *frame-sequential* method. Any interference during the transmission of any frame could cause the entire image to be useless and an improved method of *line sequential* transmission was adopted. In this case each line was sent three times, each time using a different filter. The basic waveform for these methods is shown in the figure below.



Monochrome SSTV signal

Later developments used luminance and chrominance signals instead of the usual red-green-blue signals. The first 50 to 70% of the scan line contains the luminance information which is a weighted average of the RGB signals. The remaining part of the line contains the chrominance or colour information. This choice made the method compatible with the monochrome system as it could use the first part of the line to display the monochrome image and ignore the chrominance information.

Although this encoding method reduced the time to transmit a frame from 24 s to 12 s, it produces bad quality when it encounters sharp, high contrast edges. The newer modes have therefore returned to RGB encoding.

Fast-Scan Television

Fast-scan television in the amateur service is a wideband mode that follows standard broadcast scan rates. Due to the wide bandwidth (several MHz), this mode is restricted to the UHF and microwave bands. One TV channel would consume all the space on all the HF bands up to 50 MHz!

31.5 Digital Voice—The Future?

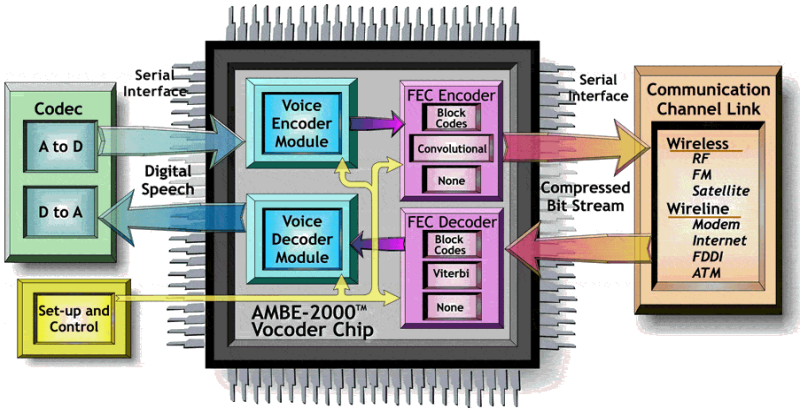
Digital modes offer advantages over their analogue counterparts and are widely used in most modern voice communication systems such as the public switched telephone network (PSTN) and cellular networks. Digital detectors need only to ascertain whether the received signal represents a digital zero or a digital one. Coding schemes have also been devised to detect and correct possible errors in the transmission making the method robust even under poor propagation conditions. The digital signals are also readily processed by advanced methods using DSP techniques to enhance their quality.

Digital speech coding can be classified into two types: *waveform coding* and *source model coding*. Waveform coding techniques involve the quantisation of the speech waveform at high rates to produce high quality speech at the cost of a high bit rate. An 8-bit A/D converter sampling at 8000 samples per second would produce a bit rate of 64 kb/s! More complex voice coders or *vocoders* use Adaptive Differential Pulse Code Modulation (ADPCM) in which prediction with differential quantisation is used to reduce the data rate to between 24 and 32 kb/s.

Model coding vocoders use a parametric model to approximate short segments of speech (between 10 and 40 ms). The speech is modeled with a fundamental frequency, a set of spectral coefficients and a set of frequency-dependent voicing decisions. This multiband voicing information and algorithms to analyse and synthesise speech have resulted in the availability of Advanced Multiband Excitation (AMBE) vocoders that can provide high quality speech at rates between 2 and 5 kb/s. These low bit rates allow the digital signal to be transmitted by HF radio without exceeding the normal 2,5 to 3,0 kHz bandwidth.

Such a system was developed for amateur radio application in 1998 by Charles Brain, G4GUO, and Andy Talbot, G4JNT. Their system is based on an AMBE1000+ speech vocoder operating at 2400 b/s plus 1200 b/s of forward error correction. Thirty-six tone carriers spaced at 62,5 Hz are used, producing an overall bandwidth of 312,5 to 2500 Hz using DQPSK modulation.

A typical application using a vocoder as in the above system is shown below:



A digital speech system for radio transmission

A digital voice system can be added to any SSB transceiver (even the older tube types) without the need for any internal modifications to the radio. The use of digital voice can provide better signals than the addition of a linear amplifier.

The use of digital voice in amateur radio is still in its infancy and ideally lends itself to experimentation and development.

The single biggest obstacle to the use of digital voice systems revolves around its behaviour when many different signals coexist in the same bandwidth, such as often happens on amateur radio. So far, digital signals do not match the discrimination ability of a human ear with analogue signals.

VoIP—Voice over Internet Protocol

Voice over Internet Protocol (VoIP) is not a radio communication mode or modulation type, and is mentioned here only as it is utilised in various networks which combine radio communication with communication via the Internet.

VoIP, also called IP Telephony, Internet telephony, Broadband telephony, Broadband Phone and Voice over Broadband is the routing of voice conversations over the Internet or through any other IP-based network.

Amateur radio has adopted VoIP by linking repeaters and users with Echolink, IRLP, D-STAR, Dingotel and EQSO. Echolink and IRLP are programs/systems based upon the Speak Freely VoIP open source software. Echolink allows users to connect to repeaters via their computer (over the Internet) rather than by using a radio. By using VoIP, amateur radio operators are able to create large repeater networks with repeaters all over the world.

Summary

Morse telegraphy can be read by ear at speeds of up to 60 words per minute. A computer can be used to assist in reading and sending Morse code.

RTTY is one of the first data communication modes that came into widespread use. FSK with a shift of 170 Hz is used, with a data rate of 45,45 Bd. RTTY is widely used to casual keyboard-to-keyboard chatting and for working DX. Because it is slow and error prone, it is seldom used for data transmission.

AMTOR is a development of RTTY. It is no longer in use, having been displaced by more sophisticated protocols and by RTTY itself.

ASCII is a code in widespread use in computer systems. It uses a 7-bit code with an extension to 8 bits for error checking.

Packet Radio uses other stations in direct radio to retransmit packets to the destination, that may not be reachable directly. Its structure is very similar to that of the Internet. Modern packet networks support standard TCP/IP protocols.

APRS provides real-time tracking of mobile or portable amateur radio stations.

PSK31 is an HF keyboard-to-keyboard protocol with good weak-signal characteristics.

WSJT is a suite of weak-signal digital modes providing huge advantages for propagation modes like EME and meteor scatter.

CLOVER is an adaptive HF system that adjusts itself to propagation conditions.

PACTOR is a combination of packet radio and AMTOR for HF, providing very low error rates at speeds of 100 to 200 b/s.

Facsimile provides high-definition images in normal speech bandwidth.

SSTV provides TV-like pictures in normal speech bandwidth, at the cost of long frame durations. Many different modes in use to provide colour pictures are in use.

FSTV uses normal broadcasting bandwidth and frame rates. Because of the bandwidth, it is useable only on UHF and above.

Digital speech may displace analogue speech signals in future, offering quality and bandwidth efficiency advantages. There is lots of room for experimentation. One of the obstacles to be overcome is the operation of digital voice systems when many signals are present simultaneously.

Modern digital communications systems often use a generic computer with a generic sound card is connected to a standard SSB transmitter. Using suitable software, digital signals are transmitted. At the receiving end, a standard SSB receiver is used to recover the signals, and a standard sound card demodulates the signals and displays them on the screen, or renders them as speech, or interprets them and shows them as a waterfall or in some other form acceptable to the operator. Digital modes are becoming extremely popular on the air.

Revision Questions

1 Morse code telegraphy:

a. Is obsolete and no longer used in amateur radio.
b. Uses wide bandwidth.
c. Requires a computer to send and receive.
d. Provides communications at up to 60 words per minute with minimal equipment.

2 RTTY is used for:

a. Sending large files.
b. Communicating very quickly.
c. Working DX stations.
d. Typing Teletubbies.

3 AMTOR:

a. Is very popular and has replaced RTTY almost completely.
b. Provides high-speed data transfer.
c. Is almost completely obsolete.
d. Is much more popular than RTTY.

4 ASCII:

a. Encodes decimal digits as separate nibbles.
b. Is easily converted into binary numbers.
c. Is widely used in computers.
d. Is used to send Morse code by sitting on it repeatedly.

5 Packet radio stations:

a. Require big antennas because they must have a wide coverage area.
b. Use nearby stations to relay messages to distant stations.
c. Require high power to achieve long distances.
d. Fit into small cardboard boxes called Packets.

6 APRS provides:

a. Position reporting using a GPS receiver.
b. Packet reporting using a TNC.
c. Propagation reports for weak signals.
d. Weak-signal protocols for EME and other specialised modes.

7 WSJT is a:

a. Modulation mode for meteor scatter.
b. Modulation mode for tropospheric scatter.
c. Modulation mode for EME.
d. Suite of programs with all of the above.

8 PACTOR provides:

a. Very high data rates.
b. Error-free data transmission on VHF.
c. Low error rates and medium throughput on HF.
d. Raw keyboard-to-keyboard teletype on HF.

9 FSTV differs from SSTV in that:

a. SSTV has much higher quality.
b. SSTV uses much more bandwidth.
c. SSTV uses far more colours.
d. SSTV takes a long time to transmit one low-quality frame.

10 Digital speech:

a. Is mature and requires no further experimentation.
b. Works well when many people are talking.
c. Offers only advantages relative to analogue speech.
d. Has a long way to go, especially when many signals are present simultaneously.

11 Digital modes:

a. Require sophisticated specialised equipment.
b. Are becoming less and less popular.
c. Have a high barrier to entry.
d. Use free software and commonly-available computer equipment.

Chapter 32: Safety Considerations

32.1 The Human Body

The human body has a number of very sensitive sensory and control systems. Most of these systems use a combination of chemical and electric mechanisms.

Some of these systems can be easily disrupted by electricity and by heat.

As little as 20 mA of current flowing through your torso can disrupt your heart, causing a fatal cardiac arrest.

There is more than just anecdotal evidence that RF fields delay healing of wounds and structural damage. RF exposure has been linked to increased risk of various cancers, including leukemia. Electricity and cellphone companies have spent huge money on disproving such links, yet concerns linger. If you want to play safe, do not expose yourself or anyone else to more RF than necessary.

32.2 Mains Power Supply

The mains power supply in South Africa is mostly 240 V at 50 Hz. Touch contact can cause lethal currents to flow. The most dangerous situation is when you touch the mains supply with a finger while there is a conduction path to ground. Concrete is a relatively good conductor of AC, so a barefoot operator on a concrete floor should not be working with electricity. Use shoes with rubber soles. Use insulated tools. Do not remove covers from equipment unless the equipment is unplugged and turned off.

When you install equipment, some points must be noted:

- Insulated wire with a suitable voltage rating must be used for all connections.
- Learn the colour code for mains wires. The local code is green/yellow for earth, brown for live and blue for neutral. However, imported equipment may have cables with different codes.
- All exposed metal surfaces (including equipment cases) must be properly earthed.
- When mains power is switched, a two-pole switch must be used to switch both the live and neutral lines.
- The plugs used must have a suitable current rating for the equipment, and mains outlets must not be overloaded by having too many plugs connected.

You should have a single “master switch” that is known to everyone who lives with you and can be used to turn off the mains supply to all your equipment. Your family can then safely disconnect the mains supply in the event that you are incapacitated by an electric shock.

32.3 High Voltages

High voltages can exist in an amateur radio station for two reasons:

- **Antenna elements can carry high voltages.** The tips of driven antennas such as dipoles and verticals are especially hazardous. Other high voltage points may include linear loading wires. The higher the power being used, the higher the voltages can go. Do not ever touch antenna elements while there is a risk of transmitting. Some medical instruments use RF to cut human flesh. Leave the surgery to the professionals!

- **Some equipment uses high voltage supplies.** Tube-type equipment can use voltages of up to 5 kV. These voltages are much more lethal than the mains supply, mostly because they can cause heating and charring that can lead to even higher

currents. Do not ever open the enclosures unless the equipment has been unplugged, the high voltage interlocks are in place and the power supply filter capacitors have been given sufficient time to discharge. There are *bleeder resistors* for that purpose, but they may be faulty.

32.4 Lightning

Lightning is of particular interest to radio amateurs, as amateurs like to live on high terrain and erect high structures. Lightning strikes are therefore more likely than for other members of the population.

Direct lightning strikes cause havoc with equipment. There is little you can do about it when there is a direct strike. However, you can take some precautions against indirect strikes, in close proximity to the station:

- Leave all mains, antenna and control cable connections unplugged when the station is not in use.
- Ground all feedlines at the base of the tower and at the entrance to the building.
- Ground all antenna structures using guidelines published by the SABS. Earth spikes and radials must be used. Buried metal water pipes may be useful, but very few modern installations are still made from water.
- Use surge arresters on all antenna and mains cables.
- Unplug all equipment when lightning is experienced close by. As sound travels at about 330 m/s, you can count the delay between lightning and thunder. At 5 s, the lightning is about 1,5 km away. It is past time to unplug!

Revision Questions

1 Where should the green wire in an AC power cord be attached in a power supply?

a. To the fuse.
b. To the hot side of the power switch.
c. To the chassis.
d. To the meter.

2 What safety feature is provided by a bleeder resistor in a power supply?

a. It improves voltage regulation.
b. It discharges the filter capacitors.
c. It removes shock hazards from the induction coils.
d. It eliminates ground-loop current.

3 For safety in any radio installation it is good practice:

a. To only use plastic piping for earthing.
b. To use unearthed metal piping.
c. Unearth all metal cases.
d. Install a master safety switch known to all in the house.

4 For safety reasons, all exposed metal work in an amateur station should be:

a. Connected to the mains neutral.
b. Free of earth connections.
c. Left completely floating.
d. Connected to a good earth.

5 When wiring up equipment:

a. Any wire available will do.
b. All plastic or insulated wires are suitable.
c. Insulated wires, suitable for the voltages, must be used.
d. Uninsulated wires are suitable.

6 Switches for breaking mains current should be:

a. Single poled and the live leads only broken.
b. Single pole low amperage switches.
c. Double poled and both live and neutral leads broken.
d. Knife switches without covers for easy access.

7 When plugs are used to connect transmitting equipment requiring high current to the mains:

a. Two-pin 5 A plugs without an earth pin are suitable.
b. 10 A three pin plugs can be used.
c. Wires can be put directly into the female plug.
d. A 16 A three pin plug should be used.

8 Radio Frequencies are used in microwave ovens for cooking purposes. In a radio station care must be taken:

a. To ensure that the power is on by touching RF points with wet fingers to feel for voltage.
b. To work on RF equipment with the covers off.
c. To adjust antennas whilst full power is applied to the antenna.
d. To screen off all RF sources from facial and bodily contact.

9 High capacitance capacitors left on work benches or other available places should be:

a. Passed to another person whose bodily contact can cause a reaction.
b. Should be stored away whilst under load.
c. Should be left lying around with impunity.
d. Should be discharged and stored.

Chapter 33: Before You Go

33.1 Meeting the Standard

T/R 61-02 Annex 6 specifies what you must know before you are let loose into the wide world of amateur radio. The syllabus is also used for the Radio Amateur Examination in South Africa. It is freely available on the Internet.

Once you have completed the study guide, you should comply with the level of knowledge specified. Please look at the document to ensure that you haven't missed anything.

The first page covers some basic operations that are not directly covered in the syllabus. Specifically, you must have a grasp of basic mathematical operations. Most of them have been used in the syllabus, but there are a few terms that are not specifically mentioned. The following list of skills comes from Annex 6:

> Candidates must know the following mathematical concepts and operations:
> − adding, subtracting, multiplying and dividing
> − fractions
> − powers of ten, exponentials, logarithms
> − squaring
> − square roots
> − inverse values
> − interpretation of linear and non-linear graphs
> − binary number system
> d) Candidates must be familiar with the formulae used in this syllabus and be able to transpose them.

33.2 Writing the RAE

The format of the examination
The examination consists of two papers:

- **The Technical paper** contains 60 questions. In principle, two hours are allocated for this paper.

- **The Regulations and Procedures paper** contains 30 questions. In principle, one hour is allocated for this paper.

You have to obtain at least 50% for each of the papers, and an average of 65% for both. A total of three hours is allowed.

Questions are graded, so that you will get a selection of easy questions and some slightly more difficult ones. If you know that your technical knowledge is a little shaky, ensure that you do well in the regulation paper. If you get 80% for regulations, you only need to get 50% on the technical paper, taking a lot of pressure off you.

Rumour has it that the regulation paper can be answered in much less than the intended hour. If you can do so, you obviously gain more time to focus on the technical paper.

The formula sheet

A formula sheet with all the most important formulas is supplied, so you don't have to memorise all the formulas. However, digging through formulas to identify the right one is a time-consuming pastime, so it is worth your while to make sure that you can quickly and easily identify the one you are looking for, and know how to rearrange and use it. Here is the formula sheet that will be provided in the exam:

$R_T = R_1 + R_2 + R_3$	$\frac{1}{R_T} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$	$V=IR$
$V_{Out} = V_{In} \frac{R_2}{R_1 + R_2}$	$P = VI = \frac{V^2}{R} = I^2 R$	$V_{RMS} = \frac{V_{Peak}}{\sqrt{2}}$
$\frac{1}{C_T} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}$	$C_T = C_1 + C_2 + C_3$	$X_C = \frac{1}{2\pi \mathrm{f} C}$
$L_T = L_1 + L_2 + L_3$	$\frac{1}{L_T} = \frac{1}{L_1} + \frac{1}{L_2} + \frac{1}{L_3}$	$X_L = 2\pi \mathrm{f} L$
$f = \frac{1}{2\pi\sqrt{LC}}$	$t = \frac{1}{f}$	$\tau = CR$
$Q = \frac{2\pi fL}{R}$ or $Q = \frac{1}{2\pi fC}$	$Q = \frac{f_C}{f_U - f_L} = \frac{centre\ frequency}{bandwidth}$	$V_S = V_P \frac{N_S}{N_P}$
$R_S = \frac{R_M}{(n-1)}$	$I_S = I_P \frac{N_P}{N_S}$	$R_P = R_L \left(\frac{N_P}{N_S} \right)^2$
$I_C = \beta I_B$	$\text{Gain (loss)} = 10\ Log_{10} \frac{\mathrm{P_{Out}}}{\mathrm{P_{In}}}$ dB	$I_{RMS} = \frac{I_{Peak}}{\sqrt{2}}$
$c = 3\,x\,10^8$ m/s	$\text{Gain (loss)} = 20\ Log_{10} \frac{\mathrm{V_{Out}}}{\mathrm{V_{In}}}$ dB	*ERP = power x gain (linear)*
$V = \mathrm{f}\ \lambda$	$\lambda = \frac{300}{f_{MHz}}$	$\lambda = \frac{c}{f} = ct$

Answering multiple-choice questions

Like anything else in life, writing multiple-choice exams requires specific skills. Even if you know everything that you need to know about the content, you still need to practice those exam-writing skills.

Enough sample questions are provided in this study guide. They are representative of the questions you will get in the RAE. If you have successfully answered all these questions, you are probably ready for the RAE. However, here are some practical hints for the exam room:

- **Plan your time:** The first thing you must do is to look at the entire paper and plan. If there are 100 questions and 180 minutes (three hours), you have a little under two minutes per question. Keep track of your progress. If you are half way through the time period and have not answered half the questions, you need to speed up. Conversely, if you are halfway through the time period and have answered more than half, you can probably afford to relax a little.

- **Relax:** It is not the end of the world if you don't answer all the questions. The pass mark is only 50% for each section and 65% in aggregate, so you can miss half the questions in one section and almost a quarter in the other, or you can miss one-third of all the questions and still pass.

- **First decide on the right answer:** Before you look at the answers provided, first decide what the right answer is. Often, looking at the answers first could lead you into a wild goose chase that may lead to a wrong answer, or at least waste lots of your time. Most often, if you know the approximate answer, it is possible to select the right answer from the options provided. It may not even be necessary to calculate the exact answer.

- **Pick the most correct answer.** Many multiple-choice examination papers will say so, but some may not. The general principle with multiple-choice questions is to mark the *most correct answer*. Sometimes you may think that none of the answers are exactly right. Pick the one that is closest to the truth. You may also think that more than one answer makes sense. Pick the one with the least uncertainties. If you calculate an answer and your answer is not one of the choices, you have probably made a mistake. Re-check your answer. If you are certain that your answer is right, pick the option that is closest to your answer.

- **Do not linger on a question:** If you cannot answer a question, mark it as unanswered and carry on. Remember that there are a few questions that are deliberately a little harder than the others. You may need more time to think about it, and it is a much better idea to first answer all the questions that you are sure of. Once you have answered the entire paper and there is some time left, re-plan your time. Count the unanswered questions and see how much time you have left. You will probably have more time left per question than you had in the beginning. You can now return to the unanswered questions and look through them again, spending a bit more time on each one.

- **Guess intelligently:** If you do not know the answer to a question, see if you can eliminate some of the options. Some options are obviously incorrect, and you can eliminate them completely from your selection. If you guess one of four answers, your likelihood of success is only 25%. If you guess one of three, your chances are 33%. If you guess one of two, the odds have risen to 50-50.

Typographic conventions

In this book, we have adhered to ISO-based conventions for writing units, including the Greek letters ß, Δ, η, μ, λ, π and Ω. You should be comfortable with all those symbols, as they are the ones you will encounter in real life.

Unfortunately, the examination system has limitations in terms of the symbols that it can accommodate, and some of those listed above are not possible to represent. The examination may therefore use some alternative notations that date from the days of letter-setting and typewriters.

As you may occasionally see these references in old magazines and perhaps even in books, learning to deal with them is not entirely a waste of time. Here are the most important deviations encountered in old documents (and possibly in the exam):

- **Resistance:** Most practical resistances can be expressed in Ω, kΩ or MΩ. You may encounter references to "ohm", "kilohm" and "megohm" respectively.

- **Wavelength:** Instead of "λ", you may see L or L_W or even WL.

- **Transistor gain:** You may see h_{fe} or h_{FE} figures instead of ß. The three are not all exactly the same thing, but they are closely related and good enough for entry-level design work.

- **Millionths:** Many practical capacitances in power supplies and audio circuits are measured in µF and some measurements are made in µm. In informal usage, a lot of people use “uF” or “MFD” for µF and “um” or “micron” for µm. Fortunately, because there is no standard SI prefix called “u”, the room for confusion is limited.

Appendix A: Glossary of Abbreviations

ß	Transistor current gain (beta, Greek small b)
Δ	Change (delta, Greek capital D)
η	Efficiency (eta, Greek small e)
µ	micro, ÷ 1 000 000 (Greek small m)
λ	wavelength (lambda, Greek small l)
π	3,141 592 653 589 793 238… (pi, Greek small p)
Ω	ohm (omega, Greek capital O)
A	Ampere
A	Index describing solar activity
AAA	All Africa Award
AC	Alternating current
ACK	Acknowledgement
ADC	Analogue to Digital Converter
AF	Audio Frequency
AGC	Automatic Gain Control
ALC	Automatic Level Control
AM	Amplitude Modulation
AMBE	Advanced Multiband Excitation (digital speech)
Amsat	Amateur Satellite Corporation
AO	Amsat Oscar (see Amsat, Oscar)
APRS	Amateur Position Reporting System
ARQ	Automatic Repeat Request
ASCII	American Standard Code for Information Interchange
ATU	Antenna Tuning Unit
b/s	bits per secon
BBS	Bulletin Board System
BCD	Binary Coded Decimal
Bd	baud (symbols per second)
BFO	Beat Frequency Oscillator
BJT	Bipolar Junction Transistor
BPF	Bandpass Filter
BPSK	Binary Phase Shift Keying
BSF	Bandstop Filter
c	centi (÷100)
C	Coulomb
C	Capacitance
CB	Common Base (amplifier)
CC	Common Collector (amplifier)
CCITT	International Telephone and Telegraph Consultative Committee
CD	Compact Disk
CE	Common Emitter (amplifier)
CEPT	European Conference of Postal and Telecommunications Administrations
CIVIL	Mnemonic to remember phase relationships
CMOS	Complementary Metal-Oxide Semiconductor
CQ	General call to any station
CRT	Cathode-Ray Tube
CW	Continuous Wave (telegraphy)
d	deci (÷10)
D	deka (x10)
DAC	Digital to Analogue Converter
dB	decibel
dBd	decibel compared to a dipole
dBi	decibel compared to isotropic
dBm	decibel compared to 1 mW
DBM	Doubly-Balanced Mixer
dBW	decibel compared to 1 W
DC	Direct Current
DC	Direct Conversion (receiver)
DDS	Direct Digital Synthesis
DFT	Discrete Fourier Transform
DSB	Double Sideband
DSB-SC	Double Sideband Suppressed Carrier
DSP	Digital Signal Processing
DX	Long-distance work (old CW abbreviation)
DXCC	DX Century Club (award for 100 countries)
E	Electric field
EF	Emitter Follower (amplifier)
EHF	Extremely High Frequency
EIRP	Effective Isotropic Radiated Power (see ERP)
EMC	Electromagnetic Compatibility
EME	Earth-Moon-Earth
EMF	Electromotive Force
ERP	Effective Radiated Power (see EIRP)
E_S	Sporadic E
f	Frequency
F	Farad
F/B	Front-to-Back ratio
FET	Field-Effect Transistor
FFT	Fast Fourier Transform

FIR	Finite Impulse Response (digital filter)
FM	Frequency Modulation
FOT	See OTF
FSK	Frequency Shift Keying
FSM	Field-Strength Meter
FSTV	Fast Scan Television
FSW	Frequency Setting Word
G	Giga (x 1 000 000 000)
GaAs	Gallium Arsenide (semiconductor)
GDO	Grid-Dip Oscillator
GMT	Greenwich Meridian Time (obsolete)
GND	Ground
GP	Groundplane
GPA	Groundplane Antenna
GPS	Global Positioning System
h	hekta (x100)
H	Henry (inductance)
H	Magnetic field
HAREC	Harmonised Amateur Radio Examination Certificate
HF	High Frequency
HPF	Highpass Filter
HT	High Tension (power supply)
HT	Handy Talkie (shack on belt)
Hz	hertz
I	Current
IARU	International Amateur Radio Union
IC	Integrated Circuit (chip)
ICASA	Independent Communications Authority of SA
IF	Intermediate Frequency
IIR	Infinite Impulse Response (digital filter)
IMD	Intermodulation Distortion
IOTA	Islands On The Air (award)
IP	Internet Protocol
IRC	International Reply Coupon
IRLP	International Repeater Linking Project
ISM	Industrial, Scientific, Military
ITA2	International Telegraph Alphabet number 2 (Baudot)
ITU	International Telecommunications Union
JT	Joe Taylor K1JT
k	kilo (x 1000)
K	K index of solar activity
K	kelvin (unit of temperature)
LC	Inductor-Capacitor network
LED	Light-Emitting Diode
LF	Low frequency
LO	Local Oscillator
LotW	Logbook of the World
LPA	Log-Periodic Array
LPDA	Log-Periodic Dipole Array
LPF	Lowpass Filter
LSB	Lower Sideband
LSB	Least Significant Bit
LUF	Lowest Useable Frequency
m	metre
m	milli (÷ 1000)
M	mega (x 1 000 000)
MFD	microfarad µF
MOS	Metal-Oxide Semiconductor
MOSFET	See MOS, FET
ms	millisecond
MS	Meteor Scatter
MSB	Most Significant Bit
MUF	Maximum Useable Frequency
MW	Medium wave
n	nano (÷ 1 000 000 000)
NBFM	Narrow-Band FM (see FM)
NiCd	Nickel-Cadmium (a battery technology)
NPN	Type of transistor (see PNP)
Oscar	Orbital Satellite Carrying Amateur Radio
OTF	Optimal Traffic Frequency
p	pico (÷ 1 000 000 000 000)
PA	Power Amplifier
PAL	Phase Alternate Line (TV standard)
PBBS	Packet Bulletin Board System (see BBS)
PC	Personal Computer
PC	Printed Circuit
PCB	Printed Circuit Board
PEP	Peak Envelope Power
PIN	P-intrinsic-N (type of diode)
PLL	Phase-Locked Loop
PM	Phase Modulation
PN	PN junction (semiconductor)
PNP	Type of transistor (see NPN)
PSK	Phase-Shift Keying
PSTN	Public Switched Telephone Network
PSU	Power Supply Unit
Q	Quality factor
QPSK	Quadrature Phase-Shift Keying
QSL	QSL Card (postcard to confirm amateur contact)
R	Resistance
RC	Resistance-Capacitance network
RF	Radio Frequency

RFC	Radio-Frequency Choke
RFI	Radio-Frequency Interference
RGB	Red, Green and Blue
RL	Resistance-Inductance network
RMS	Root-Mean-Square (effective voltage of a signal)
RST	Readability-Strength-Tone
RTTY	Radio Teletype
Rx	Receiver
s	second (unit of time)
S	siemens (unit of conductance)
SAE	Self-Addressed Envelope
SARL	South African Radio League
SASE	Self-Addressed Stamped Envelope
SAST	South African Standard Time
SDR	Software-defined radio
SFI	Solar Flux Index
SHF	Super-High Frequency
Si	Silicon
SI	French for the Metric System
SLS	Sidelobe Suppression
SMPS	Switchmode Power Supply
SN	Sunspot Number
SOTA	Summits on the Air
SSB	Single Sideband
SSTV	Slowscan Television
SWR	Standing Wave Ratio (see VSWR)
T/R	Transmit/Receive
TNC	Terminal Node Controller
TRF	Tuned Radio Frequency (receiver)
Tx	Transmitter
UHF	Ultra-High Frequency
UI	Unnumbered info (packet frame)
USA	United States of Merica
USB	Upper Sideband
UTC	Universal Coordinated Time
V	volt
V	Voltage
VCO	Voltage-Controlled Oscillator
VCXO	Voltage-Controlled Crystal Oscillator
VFO	Variable-Frequency Oscillator
VHF	Very High Frequency
VoIP	Voice over IP (see IP)
VSWR	Voltage Standing-Wave Ratio (see SWR)
VXO	Variable Crystal Oscillator
WAS	Worked All States (award)
WAZ	Worked All Zones (award)
WAZS	Worked All ZS (award)
WSJT	Weak Signal by Joe Taylor (software suite)
WSPR	Weak Signal Propagation Reporter (software suite)
X	Reactance
X_C	Capacitive Reactance
X_L	Inductive Reactance
Z	Impedance

ISBN 978-0-620-69471-1