ANALOG SEEKrets

DC to daylight

… a master class in electronics design

Leslie Green ACGI CEng MIEE

ANALOG SEEKrets

DC to daylight

Contact: logbook@lineone.net

www.logbook.freeserve.co.uk/seekrets

PUBLISHER

©

Future Science Research Press,
30 High Street, Didcot, Oxfordshire,
OX11 8EG, United Kingdom.

Printed and bound in the United Kingdom by CPI Antony Rowe, Eastbourne.

First Printing: March 2007

Second Printing: October 2007 (with minor corrections)

PDF version: March 2011

Cataloguing-in-publication data

Green, Leslie 1957 - …

Analog Seekrets, DC to daylight.

ISBN 978-0-9555064-0-6

621.3815

TK7867

SAFETY NOTICE

No part of this text shall be used as a reason for breaking established safety procedures. In the unlikely event of a conflict between this text and established safety procedures, consult the relevant responsible authority for guidance.

LEGAL NOTICE

Whilst every effort has been made to minimise errors in this text, the author and publisher specifically disclaim any liability to anyone for any errors or omissions which remain. The reader is required to exercise good judgement when applying ideas or information from this book.

DEVICE DATA

Where specific manufacturers' data has been quoted, this is purely for the purpose of specific illustration. The reader should not infer that items mentioned are being endorsed or condemned, or any shade of meaning in between.

ACKNOWLEDGEMENTS

This book would not have been possible without the assistance of over 100 personally owned text books, boxes full of published papers, and years of both assistance and debate with colleagues.

All SPICE simulations were done using SIMetrix v2.5.

Mathematical simulations and plots were done using Mathcad 13 (Mathsoft) and MS Excel 2003.

Word processing was done using MS Word 2003 on an AMD Athlon 62 X2 4200+ processor using 2GB of RAM. PDF output was produced using PDF995 by Software995.

↦

On the front cover is an F-band (90 GHz – 140 GHz) E-H tuner from Custom Microwave Inc. On the rear cover is a 2D field plot of the cross-section of a round pipe in a square conduit done using 2DField (see the author's website). The unusual RC circuit on the rear cover gives a maximum voltage gain of just over 15%. Ordinary RC circuits give less than unity voltage gain.

CONTENTS

KEY:

scope	oscilloscope	
spec	specification	
pot	potentiometer	
ptp	peak-to-peak (written as p-p in some texts)	
+ve	positive	
–ve	negative	
&c	et cetera (*and so forth*)	
trig	trigonometry / trigonometric	
≈	approximately equal to	
≡	identically equal to, typically used for definitions	
//	in parallel with, for example 10K//5K = 3K333	
j	the imaginary operator, $j \equiv i = \sqrt{-1}$	
C	a capacitor	
R	a resistor	
L	an inductor	
r.m.s	RMS	
$2 \cdot \sin(\phi)$	$2 \times \sin(\phi)$	(*a half-high dot means multiply*)
$4\pi f C$	$4 \times \pi \times f \times C$	(*adjacent terms are multiplied*)
*EX	key exercises (answers to all exercises on the WWW)	
@EX	key exercises with answers at the back of the book	
alias	bold italics means an entry exists in the glossary	
aberration	italics used for emphasis, especially for technical terms	

PREFACE

This is a book on antenna theory for those who missed out on the RF design course. It is a book on microwave theory for those who mainly use microwaves to cook TV dinners. It is a book on electromagnetic theory for those who can't confront electromagnetics textbooks. In short, it is a book for senior under-graduates or junior design engineers who want to broaden their horizons on their way to becoming 'expert' in the field of electronics design. This book is a wide-ranging exploration of many aspects of both circuit and system design.

Many texts are written so that if you open the book at a specific subject area, you can't understand what the formulas mean. Perhaps the author wrote at the beginning of the book that all logarithms are base ten. Thus, you pull a formula out of the book and get the wrong answer simply because you missed the earlier 'global' definition.

Another possible source of errors is being unsure of the rules governing the formula. Take this example: $F = \log \mathrm{t} + 1$. What is the author trying to say? Is it $F = 1 + \log_{10}(t)$, is it $F = \log_e(1+t)$, or something else? In this book, I have tried to remove all ambiguity from equations. All functions use brackets for the arguments, even when the brackets are not strictly necessary, thereby avoiding any possible confusion. See also the Key, earlier.

I have adopted this non-ambiguity rule throughout and this will inevitably mean that some of the text is "non-standard". For an electronics engineer, 10K is a 10,000 Ω resistor. The fact that a physicist would read this as 10 Kelvin is irrelevant. I believe that the removal of the degree symbol in front of Kelvin temperatures [1] is a temporary aberration, going against hundreds of years of tradition for temperature scales. Thus, I will always use 'degrees Kelvin' and the degree symbol for temperature, and hope that in the next few decades the degree symbol will reappear!

Inverse trig functions have been written using the *arc-* form, thus $\arccos(x)$ is used in preference to $\cos^{-1}(x)$. In any case, the form $\cos^{-1}(x)$ is incompatible with the form $\cos^2(x)$; to be logically consistent, $\cos^{-1}(x)$ should mean $1/\cos(x)$.

Full answers to the exercises are posted on the website.

Use the password book$owner
http://www.logbook.freeserve.co.uk/seekrets

Rather than leaving the next page blank, I managed to slip in a bit of interesting, but arguably non-essential material, which had been cut earlier…

[1] J. Terrien, 'News from the International Bureau of Weights and Measures', in *Metrologia*, 4, no. 1 (1968), pp. 41-45.

CONFORMAL MAPPING: Multiplication by *j*, the square root of minus one, can be considered as a rotation of 90° in the *complex plane*. In general, a function of a complex variable can be used to transform a given curve into another curve of a different shape. Take a complex variable, $w \equiv u + j \cdot v$, (*u* and *v* both real). Perform a mathematical operation *f(w)* on *w*, resulting in a new complex variable *z*. Then $f(w) \equiv z \equiv x + j \cdot y$, (*x* and *y* both real). *w* is transformed (mapped) onto *z*. A set of points chosen for *w* forms a curve. The function *f(w)* acting on this curve creates another curve. For clarity, draw consecutive values of *w* on one graph (*w*-plane), and the resulting values of *z* on another graph (*z*-plane).

For 'ordinary' functions such as polynomials, sin, cos, tan, arcsine, arccos, sinh, exp, log, &c, and combinations thereof, the real (*x*) and imaginary (*y*) parts of the transformed variable are related. Such ordinary functions are known as 'regular' or 'analytic'; mathematicians call them *monogenic*, and say they fulfil the Cauchy-Riemann equations:

$$\boxed{\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}}$$

The term 'conformal mapping' is used because the angle of intersection of two curves in the *w*-plane remains the same when the two curves are transformed into the *z*-plane. Consider two 'curves' in the *w*-plane, the lines *u*=*a* and *v*=*b* (*a* and *b* real constants). The angle of intersection is 90°. The resulting transformed curves also intersect at 90° in the *z*-plane. The angle of intersection is not preserved at *critical points*, which occur when $f'(w) = 0$. These critical points must therefore only occur at boundaries to the mapped region.

Further differentiation of the Cauchy-Riemann equations gives the second order partial differential equations $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$ and $\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0$. These equations are recognisable as *Laplace's equation* in 2-dimensions using rectangular co-ordinates, which are immediately applicable to the electric field pattern in a planar resistive film and the electric field pattern in a cross-section of a long uniform 3D structure such as a pipe in a conduit (see rear cover field plot).

Consider lines of constant *u* in the *w*-plane as being equipotentials; lines of constant *v* in the *w*-plane are then lines of flow (flux). Intelligent choice of the transformation function turns a line of constant *u* (*w*-plane) into a usefully shaped equipotential boundary in the *z*-plane.[2] Equipotential plots can then be obtained for otherwise intractable problems. Important field problems that have been solved by conformal mapping include ***coplanar waveguide***, microstrip and square coaxial lines. There are published collections of transformations of common functions, although complicated boundaries generally require multiple transformations. One can try various functions acting on straight lines in the *w*-plane and catalogue the results for future use, much like doing integration by tabulating derivatives. Another possibility is to take the known complicated boundary problem in the *z*-plane and transform it back to a simple solution in the *w*-plane.[3] (*Schwartz-Christoffel* transformations).

Conformal mapping allows the calculation of planar resistances having a variety of electrode configurations.[4] Considering the field pattern as a resistance, the resistance of the pattern is unchanged by the mapping operation. The resistance of the complex pattern is equal to that of the simple rectangular pattern.

$R = \frac{length}{width} \times \Omega/\square$, planar resistance. $C = \varepsilon_0 \varepsilon_r \times \frac{width}{length}$ F/m, microstrip capacitance

Far from being an obscure and obsolete mathematical method, superseded by numerical techniques, conformal mapping is still important because it produces equations of complex field patterns rather than just numerical results for any specific pattern.

[2] S. Ramo, J.R. Whinnery, and T. Van Duzer., 'Method of Conformal Transformation', in *Fields and Waves in Communication Electronics*, 3rd edn (Wiley (1965) 1994), pp. 331-349.

[3] L.V. Bewley, *Two-Dimensional Fields in Electrical Engineering* (Macmillan, 1948; repr., Dover, 1963).

[4] P.M. Hall, 'Resistance Calculations for Thin Film Patterns', in *Thin Solid Films*, 1 (1967/68), pp. 277-295.

CH1: introduction

A SEEKret is a key piece of knowledge which is not widely known, but which may be 'hidden' in old textbooks or obscure papers. It is something that you should seek-out, discover or create in order to improve your skills. Whilst others may already know this SEEKret, their knowing does not help you to fix your problems. It is necessary for you to find out these trade SEEKrets and later to create some of your own. For this reason many key papers are referenced in this book.

Recent papers can be important in their own right, or because they have references to whole chains of earlier papers. There are therefore two ways to seek out more information about a particular area: you can either pick a modern paper and work back through some of its references, or you can pick an early paper and use the *Science Citation Index* to see what subsequent papers refer back to this original.

This book is written for self-study. You are expected to be able to sit down and study the book from beginning to end without needing assistance from a tutor. Thus, all problems have solutions (on the WWW) and everything is written so that it should be accessible to even the slowest advanced student.

This is advanced material, which assumes that you have fully understood earlier courses. If not, then go back to those earlier texts and find out what you didn't get.

As you read through (your own copy of) this text, have a highlighter pen readily to hand. You will find parts that are of particular interest and you should mark them for future reference. It is essential to be able to rapidly find these key points of interest at some later time. Thus, you should also go to the index and either highlight the appropriate entry, or write an additional index entry which is meaningful to you.

Many of the exercises are marked with a *. These are essential to the flow of the text, and the answers are equally part of the text. It is important for you to follow these exercises and answers as part of the learning process; write down answers to the exercises, even when the answers appear obvious. The act of writing down even the simplest of answers makes the information 'stick' in your mind and prevents you from 'cheating'. Exercises prefixed by @ have answers at the back of the book.

I don't expect you to do well with the exercises. I actually want you to struggle with them! I want you to realise that you don't yet know this subject. You are *encouraged* to seek out information from other textbooks before looking up the answer. If you just look up the answers then you will have missed 90% of the benefit of the exercises.

Unlike standard texts, the exercises here may deliberately give too much information, too little information, or may pose unanswerable questions; just like real life in fact. The idea is that this is a 'finishing school' for designers. In many senses, it is a bridge between your academic studies and the real world. In any question where insufficient information has been given, you will be expected to state what has been left out, but also to give the most complete answer possible, taking the omission into account. Don't just give up on the question as "not supplying enough information".

Words written in ***bold italics*** are explained in the Encyclopaedic Glossary. Many important concepts are explained there, but not otherwise covered in the text. The first time a technical word or phrase is used within a section it may be emphasised by the use of italics. Such words or phrases can then be further explored on the WWW using a search engine such as Google.

Be aware that many words have two or more distinct definitions. Take a look at ***stability*** in the glossary. When you read that a particular amplifier is "stable", the statement is ambiguous. It might mean that the amplifier doesn't oscillate; or it might mean that the amplifier's gain does not shift with time; or it might mean that the DC offset of the amplifier doesn't shift with temperature.

If the text doesn't seem to make sense, check with a colleague/teacher if possible. If it still makes no sense then maybe there is an error in the book. Check the website: **http://www.logbook.freeserve.co.uk/seekrets**. If the error is not listed there, then please send it in by email. In trying to prove the book wrong, you will understand the topic better. The book itself will also be improved by detailed scrutiny and criticism.

I need to introduce some idea of costing into this book. With different currencies, and unknown annual usage rates, this is an impossible job. My solution is to quote prices in $US at the current rate. You will have to scale the values according to inflation and usage, but the relative ratios will hold to some degree.

The statement "resistors are cheaper than capacitors" is neither totally true nor wholly false. A ±1% 250 mW surface mount resistor costs about $0.01, whereas a 0.002% laboratory standard resistor can cost $1400. A surface mount capacitor currently costs about $0.04 whereas a large power factor correction capacitor costs thousands of dollars. Look in distributors' catalogues and find out how much the various parts cost.

In older text books the term "&c" is used in place of the more recent form "etc.", both being abbreviations for the Latin *et cetera*, meaning *and so forth*. The older form uses two characters rather than four and therefore deserves a place in this book. Terms such as *r.m.s.* are very long winded with all the period marks. It has become accepted in modern works to use the upper case without periods for such abbreviations. Thus terms such as RMS, using small capitals, will be used throughout this book.

For the sake of brevity, I have used the industry standard short-forms:

scope	oscilloscope
spec	specification
pot	potentiometer
ptp	peak-to-peak (written as p-p in some texts)
+ve	positive
–ve	negative
trig	trigonometry / trigonometric

I may occasionally use abbreviated notation such as ≤ ±3 μA. Clearly I should have said the magnitude is ≤ 3 μA. This 'sloppy' notation is also used extensively in industry.

CH2: an advanced class

2.1 The Printed Word

Service manuals for older equipment give a detailed description of the equipment and how it works. This does not happen nowadays. Detailed information about designs is deliberately kept secret for commercial reasons. This subject is called *intellectual property*. It is regarded as a valuable commodity that can be sold in its own right.

Universities, national metrology institutes, and commercial application departments have the opposite form of commercial pressure, however. Universities need to maintain prestige and research grants by publication in professional journals. Doctoral candidates may need to ‘get published’ as part of their PhD requirements. National metrology institutes publish research work to gain prestige and public funding. Corporate applications departments produce “app notes” to demonstrate their technical expertise and to get you to buy their products. And individual authors reveal trade SEEKrets to get you to buy their books. Publication in the electronics press is seldom done for philanthropic reasons.

It takes time, and therefore money, to develop good ways of doing or making things; these good ways are documented by drawings and procedures. It is essential that you read such documentation so you don’t “re-invent the wheel”. Your new component drawings will be company confidential and require a clause preventing the manufacturer from changing his process after prototypes have been verified. The drawing cannot specify the component down to the last detail; even something as simple as a change in the type of *conformal coating* used could ruin the part for your application.

2.2 Design Concepts

Rival equipment gives important information to a designer. It is an essential part of a designer’s job to keep up to date with what competitors are up to, *but only by legal means*.

It is quite usual to buy an example of your competitor’s product and strip it down. When theirs is the same as yours, this validates your design. When theirs is better, you naturally use anything that isn’t patented or copyrighted.

There will almost always be a previous design for anything you can think of. If you work out your new design from first principles, it will probably be a very poor solution at best. The thing you have to do is to examine what is already present and see what you can bring to the product that will enhance it in some way. You have to ask these sorts of questions:

- Can I make it less expensive?
- Can I make it smaller?
- Can I make it lighter?
- Can I make it more attractive to look at?
- Can I make it last longer?
- Can I make it more versatile?

- Can I make it more efficient?
- Can I make it more effective?
- Can I make it perform the work of several other devices as well?
- Can I make it more beneficial to the end-user ?
- Can I make it more accurate?
- Can I make it more reliable?
- Can I make it work over a wider range of temperature/humidity/pressure?
- Can I make it cheaper to service and/or calibrate?
- Can I make it as good as the rival one, but use only local labour and materials?
- Can I make it easier and cheaper to re-cycle at the end of its life?
- Can I reduce the environmental impact on end-of-life disposal?

These are general design considerations that apply to all disciplines and are not given in order of importance. A new design need only satisfy one of these points. The new design has to justify the design cost. If not, then why do the work?

From a manager's point of view, this subject is considered as *return on investment*. If the company spends $20,000 of design effort, it might expect to get $200,000 of extra sales as a result. Engineering department budgets are often between 5% and 20% of gross product sales.

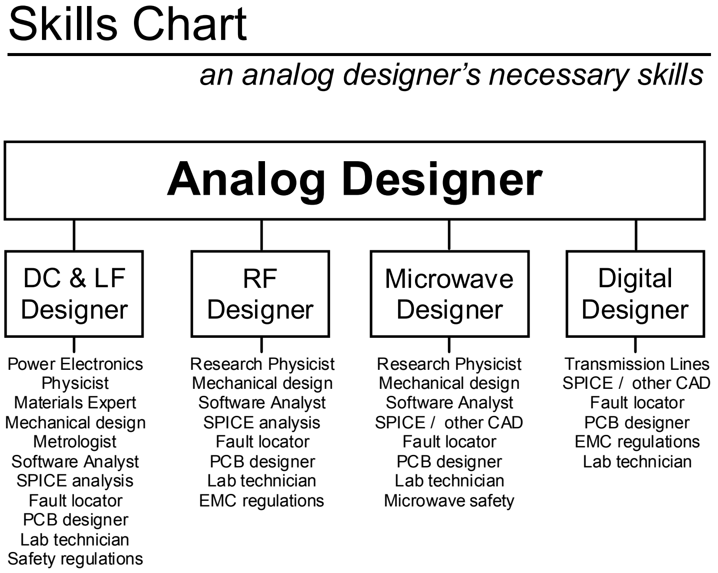
2.3 Skills

This table should not be taken too literally:

	Cheap parts	Expensive parts
poor spec	poor	unemployable
average spec	Good	average
excellent spec	GRAND MASTER	Master

Anybody can take expensive parts and make a pile of rubbish out of them. It takes skill to make something worthwhile. If you can take inexpensive parts and make a spectacular product you deserve the title Grand Master, provided that you can do it consistently, or for lots of money.

FIGURE 2.3A:


Skills Chart
an analog designer's necessary skills
Analog Designer
DC & LF Designer
RF Designer
Microwave Designer
Digital Designer
Power Electronics Physicist Materials Expert Mechanical design Metrologist Software Analyst SPICE analysis Fault locator PCB designer Lab technician Safety regulations
Research Physicist Mechanical design Software Analyst SPICE analysis Fault locator PCB designer Lab technician EMC regulations
Research Physicist Mechanical design Software Analyst SPICE / other CAD Fault locator PCB designer Lab technician Microwave safety
Transmission Lines SPICE / other CAD Fault locator PCB designer EMC regulations Lab technician

This chart suggests skills you should acquire. Your ability should at least touch upon every skill mentioned. However, I do not insist that you agree with me on this point. It is an *opinion*, not a fact. All too often people, including engineers, elevate opinions to the status of facts. Learn the facts, and decide which opinions to agree with.

I once had a boss who insisted that any wire-link modifications to production PCBs should be matched to the PCB colour to blend in. The quality department, on the other hand, wanted contrasting wire colours in order to make inspection easier. Both sides felt passionately that the other side was obviously wrong!

An opinion, theory or result stated/published by an eminent authority is not automatically correct. This is something that is not mentioned in earlier school courses. In fact, you get the *greatest* dispute amongst eminent authorities about the most advanced topics.

Having said that opinions are not facts, engineering is not the exact science you might like. Often you have to take a decision based on incomplete facts. Now, technical opinion, in the form of engineering judgement, is vital. In this case, the opinion of a seasoned professional is preferable to that of a newcomer. Furthermore, this opinion, if expressed coherently and eloquently, can carry more weight than the same opinion mumbled and jumbled in presentation. Thus, communication skills are needed to allow your ideas to be brought to fruition. Do not underestimate the importance of both verbal and written communication skills, especially in large companies or for novel products. You may need to convince technically illiterate customers, managers or venture capitalists that your idea is better than somebody else's idea. If you succeed then wealth will follow, provided your idea was good!

A related idea is that of intuition. Sometimes you can't directly measure a particular obscure fault or noise condition. In this case you may have to invent a "theory" that might explain the problem, and then test by experimentation to see if it fits. Good intuition will save a lot of time and expense. Could it be that the fluorescent lights are causing this noise effect? Turn them off and see!

Be prepared for textbooks to contain errors, both typographical and genuine author misunderstandings. And do not make the mistake of believing a theory is correct just because *everybody knows* it to be true. That a theory has been accepted for a hundred years, and re-published by many authors, does not automatically make it correct. Indeed, the greatest scientific discoveries often smash some widely held belief.

The great James Clerk Maxwell asserted in his Treatise that magnetic fields did not act directly on electric currents. A young student didn't believe this statement and tested it; the result was the ***Hall Effect***.

Experts make mistakes. Even an expert must be *willing to be corrected* in order to make them *more right* in the future. Don't hang on to a theory which doesn't work, just to show how right you were.

The difference between a professional design and a "cook book" solution or an "application note" solution is that the professional design *works*. It may not look elegant. Many professional designs require corrections, even after they have been in production for some little while. A diode here, a capacitor there. These are the vital pieces that stop the circuit from blowing up at power-off. These are the parts that fix some weird response under one set of specific conditions. These are the parts that give a factor of two improvement for little extra cost.

I am not encouraging you to need these 'afterthought' components in your designs. The point is that the difference between a workable, producible, reliable design and an unworkable 'text book solution' can be a matter of a few key components that were not mentioned in your textbooks. Don't feel ashamed to add these parts, even if the circuit then looks 'untidy'; add them because they are needed or desirable.

The best engineer is *the one who can consistently solve problems more rapidly or cost effectively than others*. Now this may mean that (s)he is just much smarter than everyone else is. Alternatively (s)he may just have a better library than others. If you have enough good textbooks at your disposal, you should be able to fix problems faster than you would otherwise do. After all, electrical and electronic engineering are mature subjects now. Most problems have been thought about before.

By seeing how somebody else solved a similar problem, you should be able to get going on a solution faster than if you have to think the whole thing through from first principles. The vital thing you have to discover is that the *techniques* of design are not tied to the components one designs with. The basics of oscillator design, for example, date back to 1920, and are the same whether you use *thermionic valves*, transistors, or superconducting organic amplifiers.

2.4 The Qualitative Statement

Engineers should try to avoid making *qualitative statements*.

Qualitative:	concerned with the nature, kind or character of something.
Quantitative:	concerned with something that is measurable.

If I say 'the resistance increases with temperature' then that is a qualitative statement. You can't use that statement to *do* anything because I have not told you how *much* it increases. If it increases by a factor of 2× every 1°C then that is quite different to changing 0.0001%/°C.

Qualitative statements come in all shapes and sizes, and sometimes they are difficult to spot. Even the term RF [Radio Frequency] is almost meaningless! Look at the table of "radio frequency bands" in the useful data section at the back of the book. Radio Frequency starts below 60 kHz, "long waves". Almost any frequency you care to think of is being transmitted by somebody and can be picked up by unscreened circuitry. The term RF therefore covers such a vast range of frequencies that it has very little meaning on its own; 60 kHz radiated signals behave quite differently to 6 GHz signals, despite them both being *electromagnetic radiation*. Light is also electromagnetic radiation, but it is easy to see that it behaves quite differently to radio frequency signals.

The original usage of *radio frequency* (RF) was for those signals which are received at an antenna {aerial} or input port, being derived from electromagnetic propagation. After the signal was down-converted to lower frequencies it was called an *intermediate frequency* (IF). The final demodulated signal was the audio frequency (nowadays base-band) signal. This idea has been extended for modern usage so that any signal higher than audio frequency could be considered as a radio frequency, although there is no reason why a signal in the audio frequency range cannot be used as a radio frequency carrier.

Because radio frequency radiation pervades the modern environment over a range extending beyond 60 kHz to 6 GHz, it is clear that all 'sensitive' amplifiers and control systems needed to be screened.

All circuitry must be considered sensitive to radiated fields unless proven otherwise by adequate testing.

Since RF covers too broad a range of frequencies to be useful for most discussions, I have tried to eliminate it from the text as much as is sensible. I like using terms such as VHF+, meaning at VHF frequencies and above, the reason being that it is *appropriately* vague. To say that a specific technique does not work above 10 MHz is unreasonable. There will, in practice, be a band of frequencies where the technique becomes progressively less workable. Thus a term like VHF (30 MHz to 300 MHz) with its ×10 band of frequencies gives a more realistic representation of where a particular phenomenon occurs. The compromise is to occasionally give an actual frequency, on the understanding that this value is necessarily somewhat uncertain.

These example qualitative statements do not communicate well to their audience:

☹ Calibrate your test equipment regularly.
☹ Zero the meter scale frequently when taking readings.
☹ Make the test connections as short as possible.
☹ Do not draw excessive current from the standard cells.
☹ Ensure that the scope has adequate bandwidth to measure the pulse risetime.
☹ Use an accurate 10 Ω resistor.
☹ Make the PCB as small as possible.

The reader will be left with questions such as: how often, how accurate, how short, how small? Answer the questions by providing the missing data. The following examples illustrate the solution to the qualitative statement problem, but d*o not consider these examples as universally true requirements*:

☺ Calibrate your test equipment at least once a week.
☺ Zero the meter scale every 5 minutes when taking readings.
☺ Make the test connections as short as possible, but certainly less than 30 cm.
☺ Do not draw more than 1 μA from the standard cells, and even then do it for not more than a few minutes every week.
☺ Ensure that the scope risetime is at least 3× shorter than the measured pulse risetime.
☺ Use a 10 Ω resistor with better than ±0.02% absolute accuracy.
☺ Make the PCB as small as possible, but certainly not larger than 10 cm on each side.

2.5 The Role of the Expert

Design engineers dislike the idea of experts. After all, every good designer thinks (s)he could have done it at least 10% faster, 10% cheaper, 10% smaller, or better in some other way. Nevertheless there are experts, defined as those people with considerable knowledge in a particular specialised field. *An* expert does not therefore have to be the very best at that particular subject. This expert could be *one* of the best.

Electronics is too large a field for one to ever truly be ‘an electronics expert’. This would be a title endowed by a newspaper. An expert would be a specialist. The term ‘expert’ is so disliked that the titles *guru* and *maven* are sometimes used to avoid it.

Finding an expert can be difficult. You will find that anybody selling their own contract or consultancy services will be “an expert” in their promotional material. Remember also that real experts may not wish to reveal their SEEKrets, particularly to a potential rival. This knowledge will have taken them years to acquire. You may therefore need to develop a rapport with such an expert before they impart some of their wisdom to you; this comes under the heading of *people skills*.

The expert should be able to predict what might happen if you do such and such operation, and if unsure should say so. The expert should be able to say how that task has been accomplished in the past and possibly suggest improvements. The expert should be able to spot obvious mistakes that make a proposal a non-starter.

What an expert should not do is stifle change and development. It is all too easy to discourage those with new ideas. The key is to let new ideas flourish whilst not wasting money investigating things which are known to have not worked.

Bad ‘Expert’: (circa 1936). An aeroplane can’t reach the speed of sound because the drag will become infinite (Prandtl–Glauert rule). The plane will disintegrate as it approaches the sound barrier.

True Expert: (circa 1936). We know that drag increases very rapidly as we approach Mach 1. We can’t test the validity of the Prandtl–Glauert rule until we push it closer to the limit. *Let’s do it*!

Given that the drag can rapidly increase by a factor of 10× as you approach the speed of sound, you should wonder if you could have been as brave as the true experts who broke the sound barrier.[1]

Every now and then an engineer should be allowed to try out an idea that seems foolish and unworkable to others. The reason is that some of these ideas may work out after all! The only proviso on this rule is that the person doing the work has to believe in it. Anyone can try a known ‘null’ experiment and have it fail. However, a committed person will struggle with the problem and possibly find a new angle of attack. In any case, young engineers who are not allowed to try different things can become frustrated, and consequently less effective.

Another form of expert is the old professor of engineering or science. It may be that this old professor is past his/her best in terms of new and original ideas. However, such an expert would know what was not known. (S)he could then guide budding PhD candidates to research these unknown areas. There is no point in getting research done into something that is already fully understood!

When struggling with a problem it is important to realise when additional help would resolve the issue more cost effectively. Don’t struggle for days / weeks when someone else can point you in the right direction in a few minutes.

[1] J.D. Anderson, *Fundamentals of Aerodynamics*, 2nd edn (McGraw-Hill, 1991).

PRE-SUPERVISOR CHECKLIST

- ❑ Is the power switched on and getting to the system (fuse intact, as evidenced by lights coming on for example)?
- ❑ Are all the power rails in spec?
- ❑ Is the noisy signal due to a faulty cable ?
- ❑ Is the system running without covers and screens so that it can be interfered with by nearby signal generators, wireless keyboards, soldering iron switching transients, mobile phones, wireless LANs &c.
- ❑ Is the signal on your scope reading incorrectly because the scope scaling is not set up for the 10:1 probe you are using, or is set for a 10:1 probe and you are using a straight cable?
- ❑ Is the ground/earth lead in the scope probe broken?
- ❑ Is the scope probe trimmed correctly for this scope *and* this scope channel?
- ❑ Is the probe 'grabber' making contact with the actual probe tip? (If you just see the fast edges of pulses this suggests a capacitive coupling due to a defective probe grabber connection.)
- ❑ Are you actually viewing the equipment *and* channel into which the probe/cable is connected?
- ❑ Is your test equipment running in the correct mode? (An ***FFT*** of a pure sinusoid shows spurious harmonic distortion if peak detection is selected on the scope. Spectrum analysers give low readings when the auto-coupling between resolution bandwidth, video bandwidth and sweep speed is disabled.)
- ❑ Has any equipment been added or moved recently? [In one case an ultra-pure signal generator moved on top of a computer monitor produced unpleasant intermodulation products that caused a harmonic distortion test to fail.]

Often, in explaining the problem to somebody else, you suddenly realise your own mistake before they can even open their mouth! Failing that, another person may quickly see your mistake and get the work progressing again quickly. That doesn't make this other person cleverer than you; they are just looking at the problem from a fresh perspective.

The difficulty comes when you are *really* stuck. You have been through the pre-supervisor list and called the supervisor in, and other colleagues, and no-one has helped. If you call in a consultant from outside you will have to pay real money. Before doing this there is the pre-consultant checklist.

PRE-CONSULTANT CHECKLIST

❑ Have you used the pre-supervisor checklist and discussed the problem "internally" with as many relevant personnel as possible?

❑ Have you established which part of the system is causing the problem? (Remove parts completely where possible.) Use a linear bench power supply to replace a suspect switched-mode supply. Reposition sub-assemblies on long leads to test the effect.

❑ Have you faced up to difficult / expensive / unpleasant tests which could settle the problem? It may be that the only way to be sure of the source of a problem is to damage the equipment by drilling holes or cutting slots. There is no point calling in a consultant to tell you what you already knew, but were unwilling to face up to.

❑ Have you prepared a package of circuit diagrams, component layouts, specs, drawings and so forth, so that the consultant gets up to speed quickly? Consultant's time is expensive, so you must use it efficiently. Also, in preparing the package you may spot the error yourself because you are having to get everything ready for somebody new.

❑ Have you tried to 'step back' from the problem and see how it would appear to a consultant? Have you thought what a consultant might say in this situation? For noise the obvious answers are board re-layout, shielding, more layers on the PCB and so forth. (If the advice is a re-layout, it is difficult to tell the difference between a good consultant and a bad consultant.) You need to ask if it is possible to model the improvements before scrapping the existing design / layout. You should get the consultant to verify the detail of the new layout, since you got the layout wrong last time.

❑ Does the consultant have proven experience in this particular area?

❑ Have you set a budget in time / money for the consultancy: a day, a week, a month &c? Consultancy can easily drag on past an optimum point as there is always more to do.

Let's suppose your company designs and manufactures phase displacement anomalisers.[†] Your company is the only one in the country that does so. You will not find a consultant with direct experience of the subject, but you may find one with expertise in the particular area you are having trouble with. The consultant will want to ask questions, and it is essential that there is always somebody on hand who is willing, able, and competent to immediately answer these questions. These questions may seem silly, but they are key to solving the problem. The problem has to be reduced to a

[†] Invented product type.

simplicity. The car won't start: is it a fuel problem, an electrical problem, or a mechanical problem?

Any problem can seem overwhelming when you are immersed in it too deeply and you are under time pressure to solve it. The consultant is not under such pressure. If the power supply is a pile of rubbish, (s)he can say so without guilt because (s)he didn't design it. (S)he can therefore be more objective as (s)he has no vested interest in protecting his/her design. You, on the other hand, may not want to 'blame' your cherished power supply as being the guilty party!

This is the point where you need to try to be objective about your own design. Come back in the next day without a tie on, or park in a different place, or get out of bed the wrong side, or something, so that you can assume a different viewpoint. You may need to look at the problem from a different angle in order to solve it.

In solving difficult problems in the past, I have always found it better to have at least two people working on the problem. It is more efficient, as there is less stress on the individual. I don't mean that there is one person being helped by another. There are two people assigned to the problem, both of whom are equally responsible for getting a solution. Neither is then 'asking for help' as such. They have been assigned to the problem by a supervisor, or project leader, in order to efficiently resolve the problem.

Often you need at least three people to get to grips with a problem. Modern systems are so complicated that it can be impossible to immediately say that the problem is due to the software, the digital hardware, or the analog hardware. Thus at least one representative of each discipline is needed. It can, and has, taken hours or days to finally establish that the problem is specifically due to one of these three sources. *Team-working skills are therefore essential* and it is also essential to be able to understand enough about the other person's field to realise when they are talking rubbish!

2.6 The Professional

The difference between a professional and an amateur is very significant. One cannot say that a professional is always better than an amateur in some specific area, or that the professional always does a better job than the amateur in some particular respect. The difference is that the professional does it for a living!

This may sound obvious, but it is a key difference. Car enthusiasts will polish the outer surface of an engine and make it spotlessly clean. They will lovingly make the wiring looms perfectly straight and generally fiddle with parts that don't really need attention. On the other hand, if you put your car into the garage to be fixed a (good) mechanic will focus on the essentials and give it back fixed in 1/10th the time an amateur might 'play' with it.

The amateur might look at the job done and shake their head, saying what a poor quality job the garage had done. The good mechanic would have done a *professional* job. The necessary amount of work to keep the car running at minimal cost. The customer would not have been pleased with a 10× larger bill!

Design effort costs money and that has to be spent wisely. It is no good spending two weeks saving $0.01 per unit if you are only ever going to make 100 units. Suppose the total design cost can be reduced by $100,000 by using an 'inferior' design that requires more labour to set up each unit. If this increases the total production cost by only $10,000 you should go for the 'inferior' design.

The design cost per hour is not your salary; it is the total cost to your employer, including heating, lighting, sickness pay &c. This total will probably be more than double your salary. To summarise, if you spend 3 hours design time at $100/hour to save $0.12 per unit on a total lifetime build of 1000 units, you have not done a good job.

***EX 2.6.1:** Is it worthwhile saving two (37 hour) weeks of design effort at $100/hour by adding a 'hand tuning' stage to the production process which adds $1.13 to the unit product cost on a total run of 5000 units?

The theory so far is incomplete because I have talked about a 'total production run' of units. If you design a particular circuit block of a finished unit, that block may be copied from design to design for many more units than this current job. For example, if you have already designed a power supply overload protection circuit, you may well decide to use this same circuit on your recently invented quantum transmogrifier[2] to get the product to market more rapidly.

It is both expensive and time consuming to produce completely new designs, so it is important to be able to re-use parts of existing designs. In the software world this would be equivalent to re-using a subroutine {function; procedure; C++ class}. This re-use can make the number of units difficult to calculate; you may have to consider future products which are not yet even planned!

Another (potentially larger) factor relates to either lost sales or gained sales; the 'lost opportunity' cost. If you are producing a unit which no-one else produces then you have 'cornered the market'. This is a very unlikely situation. In this case, if you take an extra few weeks to bring the product to market, you will not lose any sales. All potential customers will still buy your product. Even then, you may well find that the rate of people buying the product is reasonably fixed at a certain number per week. You might then decide that by being late to market you had lost potential sales. This could amount to a very substantial amount of money.

In the more usual situation, even if your product is 10× better than the competition for the same cost, you will find that being late to market with your poly-phasic bread de-humidifier[3] loses you money. Your competitor has already sold product to your potential customers.

For this reason it is usual to 'launch' products before they are completed. This doesn't just mean before the production process is completed, it is sometimes before the design process is completed. Launching a few months early is not at all unusual, but this depends on the nature of the product. Military equipment is shown to the market as much as several years before it is ready, simply because of the long delays in getting budgets approved. More expensive products are typically launched further ahead than less expensive products.

For products like integrated circuits, you may find the product launched before the silicon is finished. If you try designing this part into your new product, you can be disappointed when the silicon is a year or so late, or indeed if it is *never* produced because they just can't make it work! (This has happened.)

[2] Invented product type; transmogrify means to change in form or appearance.
[3] Invented product type; an up-market toaster for bread.

EX 2.6.2: If you hire a thrudge-o-matic[4] pulse generator for three weeks at the extortionate cost of $10,000/week (you are still on a 37 hour week, costing $100/hr) you feel you can get the design job done 2 weeks early. Marketing are claiming that you are costing the company 100 sales per week by being so late, and you know that there is a net profit on each unit of $120. Should you try to convince your boss to sign the purchase requisition for the equipment hire?

There is another important difference between the amateur and the professional and that relates to *control*. An amateur may not be in control of a circuit or a production process. Everything else seems to be 'controlling' the production process; the weather, the component supplier, the time of day, and any other randomising factor you can think of. This is not good enough for the professional. When you have to make a living at electronics, it is no good complaining that you can't make toasters today because it is too damp!

In any advanced electronics course, you should be exposed to the idea of tolerancing your circuits. This means making sure that component-to-component variations and environmental factors do not make your design fail. That's pretty important but can be overstressed. If you make lots of units and the design is not sufficiently well toleranced then you will get a lot of non-working units. Fair enough. However, you must first test the circuit and make sure it works. There is no point in tolerancing a circuit for a week, only to find out that the basic circuit doesn't work reliably for some other reason. Thus there is an interaction between the overall circuit design concept and the detailed tolerancing. You may need to do this iteratively, in the sense of sketching a circuit and seeing if the components required are sensible before refining the design.

Most of the time you can get a bigger component or a more precise component if your tolerancing exercise shows that there is a problem with power dissipation or accuracy, but don't waste your valuable time on a circuit that doesn't work repeatedly in the first place. You will find that you are automatically doing design work as you make up a prototype. What power rating of resistor goes here, what speed of opamp goes there &c? Build a quick model or simulate it to make sure the idea is sound before spending days checking its exact stability factor.

There is also another helpful indicator when building a prototype. If the circuit works easily and quickly with almost any values, then it will probably be an easy circuit to produce commercially. If it is a real nightmare to get working on the bench at the prototype stage, you should definitely consider another approach, as it is likely to be a nightmare in production as well.

You will (hopefully) have a great many ideas that you can *profitably* exchange for money or goods. It is no good being 'brilliant' and poor. If you are academically brilliant, but cannot make money from this skill, you should realise that you need to improve your ability to make money. It is not "clever" to earn less than the guy who serves hamburgers at the local fast food restaurant.

Consider the tragic case of Oliver Heaviside. You may have heard of the Kennelly-Heaviside layer in the atmosphere, the Heaviside Operator in calculus, the Heaviside step function and so forth. Perhaps you are also aware of Heaviside's contribution to long distance telephony by his accurate analysis of transmission lines. Heaviside's

[4] Invented trade name; 'thrudge' being a slang term for an unwanted spike on a displayed waveform.

contributions to electrical engineering were huge. He received acclaim from the greats of the time, medals, honours and so forth. Nevertheless he lived his whole life as a poor man and even had the gas supply to his home cutoff for an extended period due to non-payment of bills.[5] This is a salutary lesson: honours, medals and acclaim *do not pay bills*. Earn lots of money and drive expensive cars, live well, or give it away to charity, according to your own choice … but don't live in poverty because you are "too clever" to earn a living!

2.7 Reliability

Reliability is the continued working of a component or system within its specification for an extended period of time. Standard texts on the subject draw pictures of the 'bathtub diagram' showing an initial period of high failure rate, a long period of constant random failure, then a further increasing rate at the end of the component's useful life (wear-out failure).

That is the accepted wisdom of the mathematician. The components are going to fail randomly, and apart from weeding out the ones which are going to fail early using ***burn-in*** or ***stress screening***, there is nothing you can do about it. This is not acceptable for a professional; a professional has to be in control!

There are design techniques to allow a system to continue to work when one or more components fail. These are multiply redundant systems, also known as *fault tolerant* systems. For example you could have two power supplies in a system. Maybe you arrange for each power supply to be running at 1/3rd of its maximum rated power. You connect them in parallel, sharing the load equally in some manner. If one of them fails, the other takes up the load and is then run at 2/3rd of its maximum rating. This is fine so long as the power supply supervisor circuit can inform the operator that a failure has occurred, and that the faulty power supply does not fail short-circuit. These are technical problems to solve, but they are not insurmountable. The parallel redundant scheme has made the system more reliable. This argument can be extended to having several layers of series-parallel devices to get a heavily fault tolerant system.

I am more interested in the components at the moment. Why should a component fail *at all*? There are too many electronic systems in the everyday person's modern life. It is just not acceptable to be having to fix a broken electronic device of one sort or another every few days. What is needed is a strategy for producing components that have a reasonable operating period, which I would measure in tens of years. After all, buildings have been successfully designed to last for hundreds of years.

I am thinking more of the mature device, one whose function is not going to change significantly over a few decades. A service life of 20 to 100 years should be a reasonable expectation. For this you need good components. Things like light dimmers, street light controllers and hand-held DMMs fall within this category.

Let's consider a resistor made from a thin film of conducting material on a ceramic substrate. Why should it fail, ever? The statistics say that it will fail at some random time which you have no control over, but why?

[5] O. Heaviside, 'Preface, by Sir Edmund Whittaker', in *Electro-Magnetic Theory Vol I*, 3rd edn (New York: Chelsea Publishing Company, [1893] 1971).

There are four primary reasons for failure:

a) The film or substrate had a crack, blemish or some other fault when it was made. This weak spot grows with time and the component fails.
b) The material of the film or the substrate chemically decomposes with time.
c) The application or removal of power, or external ambient changes, cause the component to heat up and cool down. This amounts to a *fatigue failure* and will give the component a finite life in terms of cycles of operation, rather than an absolute time.
d) The component is over-stressed in its operating environment. It may fail on a peak overload.

When you look at the component reliability in this way you can draw some interesting conclusions. A failure of type (a) could happen at any random time. The component is inherently unreliable because *it was faulty when manufactured*, but this was not detected. All the various burn-in stress screening techniques may not cause it to fail and you are left with a component which is sitting there, waiting to fail. Not a very satisfactory situation.

If you were making struts to hold the wheels on a car and you never looked at the parts that were made, you only tested them at some compressive load, faulty ones would get through. Perhaps they were cast badly and had voids visible on the outside. These voids would provide high stress points for subsequent fatigue failure. Nobody would tolerate such a situation.

And yet with resistors, capacitors, and semiconductors, this exact same situation is tolerated. Semiconductors are in fact notorious for it. The manufacturers know that there are random defects in the wafers. They know that the bigger the silicon die {the plural is *dice*}, the more chance there is of a defect within the die, and that therefore the larger the die, the lower the yield. The solution: 100% test the devices for functionality. But if the aluminisation interconnect happens to have a defect that nearly cuts right through it, but not quite, then the die will pass this test. Even if over-stressed, the track may survive. But over the course of the next few days, months or years, this component is going to fail; long before the others in the batch.

Visual inspection at the microscopic level is required. Human inspection of small parts by microscope is possible, but expensive on something like a transistor. It would be completely infeasible on a large scale integration device like an integrated circuit. Machine inspection is the only possible solution and there is now sufficient computational power available at a reasonable cost to automatically inspect resistors, capacitors and semiconductors; defective devices can then be discarded before they are packaged. This method has a major impact on component reliability and on finished product lifetimes.

Type (b) failures, component parts chemically decomposing, are very much more difficult to handle. In order to know for certain that a material will survive for 70 years, you ideally need to have seen a 70 year old component made the same way. This takes a long time to test on a new process! There are such things as *accelerated life tests*, but they are always open to question. Does the accelerated test correctly emulate the real lifetime of the component?

The type (c) fatigue failure is very much easier to deal with than the type (b).

Accelerated life tests will find problems with the design of the component in a very predictable and quantifiable manner. Note that fatigue failure is easy to fix when you know not to concentrate stress in the component. Unfortunately current methodology does exactly this when trimming resistors. If you simulate putting a laser cut across a resistive film, the voltage stress gradient around the cut can be 10× greater than elsewhere in the film. This is a *very* poor way of trimming a resistor to attain long service life.

The type (d) failure, in-circuit over-stress, is down to the designer to fix, using the right component and circuit structure to make the design reliable.

My (qualitative) view on reliability:

- Well made components do not fail randomly.
- A component which is stressed more heavily will fail more quickly.
- Components which are under a lot of stress should be identified and derated so that they last a long time.
- Major problems with reliability occur in practice, not because of the number of components involved, but because of defective *batches* of components. (Hence batch codes on parts are a vital production fault-finding aid.)
- Visual, X-ray or ultra-sonic inspection of components and soldered joints is essential to ensuring ultimate levels of reliability.

Ceramic capacitors can be checked for micro-defects before they are put into service. An acoustic micro-imaging process using ultrasound is being done on a million ceramic capacitors a year for high reliability applications.[6] By removing the capacitors with inherent voids, delaminations {separations between the layers in a multilayer capacitor} and cracks, a higher reliability part is achieved. Also, the manufacturer gets feedback on his process and can therefore adjust it to optimise the quality of the finished parts.

I have slightly oversimplified the subject of reliability. The type (a) “growing defect” failure I have described is complicated by being mixed with type (c) fatigue failure. It is known, for example, that slight cracks in optical fibres grow with time and will eventually cause failure. It may not be possible to produce a material that is sufficiently free of micro-defects for it to never fail.

In the case of electro-migration {positive ion-migration} in integrated circuits, it is known that high current densities cause the tracks to erode and fail. The current density increases in areas with defects and the electro-migration is therefore worse at these points. You can therefore get a finite life in the conductor due to the track being eaten away by the current flow through it.[7]

In electro-migration the movement of the atoms is attributed to an “electron wind”. For Aluminium it has been shown that provided the current density is below a critical threshold of 1600 A/mm^2, electro-migration does not occur. Above the critical threshold, the rate of migration is proportional to the increase of current density.[8]

Don’t worry about electro-migration for ordinary PCB tracks. The critical current

[6] R. Carbone, and T. Adams, 'Chip Cap Flaws Investigated', in *Electronics World*, May 2001, pp. 330-332.

[7] M. Ohring, *Reliability and Failure of Electronic Materials and Devices* (Academic Press, 1998).

[8] P.-C. Wang, G. S. Cargill III, I. C. Noyan, and C.-K. Hu, 'Electromigration-Induced Stress in Aluminum Conductor Lines Measured by X-Ray Microdiffraction', in *Applied Physics Letters*, no. 72 (1998), p. 1296.

density for copper is around 1000 A/mm². This current density is not even approached on PCBs, a more usual maximum current density being 2 A/mm².

Calculating current density on PCBs is quite an exercise due to the units. PCB designers use track widths in thousandths of an inch and copper thickness in ounces. "One ounce copper" is the most usual thickness, meaning 1 ounce of copper per square foot. 1 square foot is 0.09290 m², 1 ounce is 0.02835 kg, and the density of copper is 8900 kg/m³. This gives the thickness of 1 ounce copper as $\dfrac{0.02835}{8900 \times 0.09290} = 34.3\ \mu\text{m}$.

EX 2.7.1: What is the current density on a 10 thou {0.010 inch} track carrying 50 mA on a 1 oz copper PCB?

One useful result of reliability theory is the idea that if you make PCBs with too many components on them you will get to the point in manufacturing where you are almost guaranteed to get nearly 100% faulty boards coming from the assembly line.

EX 2.7.2: Board A has 1000 components and 3000 solder joints. Board B has 70 components and 300 solder joints. Neglecting faulty components, and assuming that, on average, one solder joint in every 10,000 is open-circuit:

a) What is the probability of a board being faulty if it is of type A?
b) What if it is type B?
c) What happens to these figures if the solder joint quality drops to one faulty in every 2000 joints?

Assume that even one open-circuit joint makes a board faulty.

2.8 Engineering Theory

The purpose of theory in engineering is to:

- ☺ Reduce design time and design cost.
- ☺ Reduce the number of experiments required to produce a workable design.
- ☺ Eliminate obviously infeasible solutions which are “blind alleys”.
- ☺ Find new solutions to problems that would not be found by chance alone.
- ☺ Generalise experimental results so that, rather than remembering the results of a large number of experiments, one can predict the result of a new experiment more easily.
- ☺ Interpolate and extrapolate the measurements of real systems, allowing a calculable margin of safety, rather than requiring testing beyond the normal ratings.

Bad theories have one or more of the following characteristics:

- ☹ They do not predict new results.
- ☹ They conflict with valid experimental data.
- ☹ They unnecessarily create particles, motions or interactions which are not measurable.
- ☹ They discourage experimental work which could make new discoveries.
- ☹ They slow down further advances in the subject.
- ☹ They do the opposite of any of the valid purposes stated above such as increasing design time.

A theory does not necessarily have to cover all situations. It is possible for a particular theory to still be useful if it has a limited range of applicability. Typically, useful theories are approximations to unreasonably complicated analytically intractable general cases.

Quotation:[9]

In engineering, most problems are not amenable to exact solutions. Therefore the ability to make approximations can spell the difference between success and failure in the solution of the problem.

[9] E.C. Jordan, and K.G. Balmain, *Electromagnetic Waves and Radiating Systems*, 2nd edn (Prentice-Hall, 1968), p. 588.

CH3: tolerancing

3.1 Selection Tolerance & Preferred Values

Tolerancing is an important subject to those designing 100+ units of anything. It is not a very important subject for those designing only 1 to 5 units, however. Throughout this chapter it is assumed that the design under consideration falls into the 100+ units category.

When a manufacturer makes a component, the spread of values may be greater than is acceptable to the customer. In this case the manufacturer may sort the components into separate bins {containers}, giving a reduced spread of values in any one bin. This process is known as *selection* and the spread of values in any bin is known as the *selection tolerance*.

A resistor manufacturer could, in principle, select any resistor down to say ±0.02% accuracy. However, this would be a pointless exercise if the resistor drifted by 2% in its first month, or if it changed by 1% when the temperature changed by 10°C. Thus good engineering judgment requires that one should not select components more tightly than is justified on the basis of their stability.

When a designer works out the value of a component, the value will inevitably be some ‘floating point’ value. If it is a resistor, maybe the calculation suggests 1023.3 Ω. Typically the designer then picks an ‘off the shelf’ part, a *preferred value* close to the calculated value. Distributors stock these preferred values, making the component cheaply and immediately available.

Continuing with resistors, for simplicity, the distributor needs to stock resistors in the range of 10 Ω to 1 MΩ for general use. By using a fixed ratio between resistor values, rather than a fixed resistance, a resistor will always be available within a certain percentage of the calculated value. The scheme that has been adopted is to use simple two or three digit values, repeating every decade. Thus if the preferred value of 18 is chosen, then 180, 1800, &c will also be chosen.

If 6 values are chosen per decade, this gives a table of preferred values:

$$\begin{array}{llll} 10 & 10 \times f & 10 \times f^2 & 10 \times f^3 \\ 10 \times f^4 & 10 \times f^5 & 10 \times f^6 = 100 & \end{array}$$

With $f = \sqrt[6]{10} = 1.468$. The whole series can be evaluated by taking the 6th root of 10^N, for N between 0 and 5.

The actual number of preferred values chosen per decade is 24 and is known as the E24 series. The E12 series is then found by skipping alternate values in the E24 series. Likewise for the E6 series. For the E24 series, $f = \sqrt[24]{10} = 1.101$, and it should be clear that resistors with a ±5% spread of values will fill the range. That is the theory behind the values of resistors, capacitors and inductors; the so called ‘preferred values’ of components.[1]

E24 values only cover the ±5% range of resistors. For ±1% resistors the E96 can be used, based on the same sort of formula, but this time with the rounding done to 3 significant figures, skipping alternate values from the E192 series.

For resistors of ±0.1% and tighter tolerances you will find that they are specified as E96 values, as exact user specified values, or as simple integer values such as 1K, 2K, 5K and 10K. Parts with better than ±0.1% tolerance will not be well stocked by distributors anyway, so you will probably just specify the exact value that you require and they will be made specifically for you.

You will be able to buy resistors of very high accuracy [better than ±0.01%] as stock items, but they will have a very restricted range of values [for example 10 values only]. Ratio-matched pairs of excellent accuracy [±0.001%] will also be available as standard parts with common ratios such as 2:1, 5:1, and 10:1.

For *any* new component that you intend to use, first check its cost and availability. It doesn't matter that the part is a "standard part" from a manufacturer's catalogue. The fact that it is in their catalogue does not mean that they do make it, or even that they ever have made it. It can simply mean that they expect to be able to make it! You can really get into trouble on a project timescale by specifying a part for which samples are readily available, but for which production quantities are on 14 week *lead-times*. Long lead-times can occur with shielded inductors, switched-mode transformers, shielded connectors, specialised resistors, any microwave components, high voltage capacitors, physically large electrolytics … the list is endless. Watch out for long lead-times on these "standard items" as you might not be expecting trouble. Play safe: check price, minimum order quantity (MOQ), and lead-time before you even test the part in a prototype.

3.2 Drift

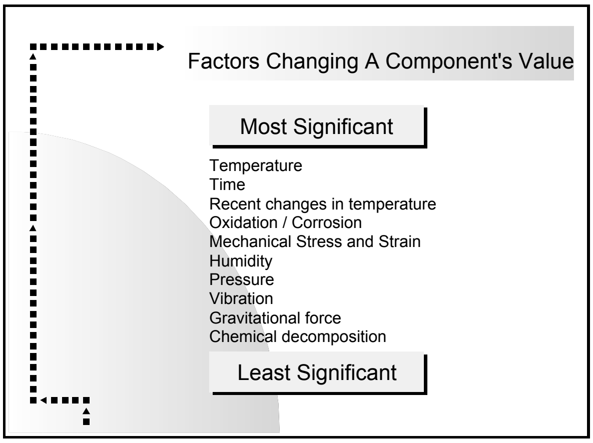
When you buy a component of a particular value, a *nominal* value or a marked value, you could say that it is guaranteed not to be that *exact* value! Given that measured values of resistance, length, voltage &c are not integers, you must expect that there will be a range of possible values for any component that you buy.

If you measure the component one day, and again the next, it *will* have a different value. Hopefully this difference will be small enough to neglect, but you must be aware of the possibility. The difference may also be smaller than the resolution of your measuring equipment, particularly if you happen to be making measurements with a 3½ digit DMM.

Here is a brief list of the some of the factors affecting a component's value. There is no need to state what type of component it is at this stage. These factors will affect all components to a greater or lesser degree. If you become involved in 'high accuracy' work you will need to take more of these factors into account.

[1] 'Preferred Number Series for Resistors and Capacitors', IEC 60063; 2nd edn (International Electrotechnical Commission, 1963), formerly IEC 63.

FIGURE 3.2A:



I have graded these for general purpose components in land-based applications. If your application is in extremes of pressure (deep sea or space) or extremes of gravitational force (aerospace) then those factors could become highly significant. Be aware that they have an effect, quantify the effect, then ignore it or take it into account.

You should note that these factors *have always been present and will always be present in the future*. Better technology will reduce the absolute amounts of the effects, but then greater accuracies will be demanded, so the factors will still need to be considered.

As a general rule, if you are working with ±5% tolerance (or greater) components then you should be able to neglect everything except the temperature and time effects. If you are working with ±0.1% tolerance (or less) components then you should definitely consider the recent changes in temperature and the humidity. Below ±0.001%, which is often expressed as ±10 ***ppm***, you should consider the whole list.

You now need to assess what effect an individual component has on the overall system spec. Perhaps only 10% of the components in a system have a significant effect on the final spec. All the components have to work, but they don't directly affect the spec. The key thing here is to look over the design and spot the most important components.

Suppose the system is a frequency counter. The user spec would include a signal sensitivity at the input, a maximum operating frequency, and an input protection voltage. How many components affect the input protection voltage? Probably somewhere between 5 and 20, and all in the input circuit area. Again signal sensitivity would be affected by another set of components in the same area, this new set inevitably including some or all of the components relating to input protection.

Maximum operating frequency would be set, or perhaps limited, by yet another specific set of components. Thus the area that needs to be looked at is not every component for every spec point. Very quickly one can home-in on the critical parts and see what effect each one has.

Keeping with the example of the frequency counter, let's look at the internal power supply requirements. How do you know how much current is going to be taken from the power supply? The answer is that you can't know for sure before you have built one. You have to make estimates of the current that could be drawn. This estimate is particularly difficult for logic and memory type devices because often, particularly for MOS and CMOS devices, the current drawn is almost proportional to the amount of digital activity and to the capacitive loading. Taking worst case numbers for the load gives an overly pessimistic power requirement which can easily be wrong by a factor of four or more.

3.3 Engineering Compromises

Anybody can come up with a “safe solution”. Add up all the worst case values and add 50% as a safety margin. If you are working on ‘life support’ or ‘system critical’ equipment you may well need to do this. For ordinary commercial designs, however, you cannot afford to do this. The cost will be too high, the weight will be too heavy, and the size will be too large. Your design will not look good compared to your competitors. Your company will not be profitable and you will be to blame.

Your design needs to be an *engineering compromise*, having the following attributes (not given in order of importance):

☺ low material cost
☺ low part count
☺ low labour cost to build and test
☺ reliable
☺ easy and cheap to service when a fault does occur
☺ quick and easy to calibrate (if necessary)
☺ minimal use of difficult-to-obtain specialist components
☺ minimal need for recalibration at fixed intervals
☺ minimal design time
☺ minimal failures during the guarantee period
☺ freedom from ***cascade failures*** that destroy the equipment and give the company a bad name.

It is impossible to put these in order of importance, this order being dependent on the nature of the product and end-user expectations. Realise that your design can never achieve the lowest values for all of the above points simultaneously. You need to expend more design time to minimise the material and labour costs. You may need to use more components, or more expensive components, to minimise failures during the guarantee period.

Failures outside of the guarantee period are not to be encouraged, since they affect the company’s reputation. However, failures during the guarantee period are directly expensive to the company and also highly visible to accountants.

If you can gain lots of ‘safety margin’ for little cost then do so. For example, if the worst measured supply current on one unit is 300 mA then you would do well to put in a 1 A regulator and have lots of margin. However, if you measured 900 mA, then a 1 A regulator is uncomfortably close to the limit. On the other hand, the next standard size up may well be a 3 A regulator. This may be significantly more expensive and bulky. Which ever way you choose, somebody will probably find fault with your decision! It is either too close to the limit or ‘over-designed’ {too safe and too expensive}.

How should you decide which way to go? First look at temperature. Does the current increase or decrease with temperature? Heat the circuit with a (domestic) hairdryer and find out. Cool the circuit with a can of freezer spray as well to make sure. Measure the current at the worst temperature.

Have you tested the circuit with the worst user stimulus? A frequency counter will probably draw more current when it is measuring at its highest frequency. Deliberately exceed the spec. It is no use designing a frequency counter which ‘blows up’ when the input frequency exceeds the maximum guaranteed operating frequency.

More important than these is to realise how the load current is defined. Investigate the circuit to see what factors affect the current. If it is due to lots of low power analog circuitry, then there is a good statistical distribution going on. The spread of supply current will be relatively low. If, on the other hand, the bulk of the current is due to a few MOS/CMOS type devices, beware! These devices take virtually no static current, drawing current only when being clocked. Hence you have to test the circuitry in many different operating modes to get the worst case. Then you have the problem of different manufacturers. 'Compatible' devices from different manufacturers can draw vastly different dynamic currents. Even from batch to batch with the same manufacturer you should expect to see at least ±15% load current variation under the same operating conditions.

If you happened to be 'unlucky' and measured a unit with low current devices then during production the maximum might go up to 30% higher than you initially measured. This would only occur where one device dominates the current load or where there is a large group of identical devices dominating the current load. If they are from the same logic family, but different type numbers, then you are back to the statistical tolerance situation and it is much safer.

3.4 Combining Tolerances

If you have cascaded attenuators then the worst case gain is found by adding the individual tolerances. (See the Appendix: Why are tolerances added?)

Mathematicians say that if you combine several *statistically independent* variables, quantities which are not related to each other in any predictable way {un-correlated}, then regardless of the probability distribution of these variables, the overall result is approximately a Gaussian (Normal) distribution. This is called the ***Central Limit Theorem***.

Mathematicians evaluate the statistical effect by adding *variances*. Electronics Engineers combine the peak tolerances in a Root of the Sum of the Squares (RSS) manner, which amounts to the same thing.

@EX 3.4.1: You have 4 cascaded attenuators with worst case gain tolerances of 1%, 2%, 1% and 3% respectively.

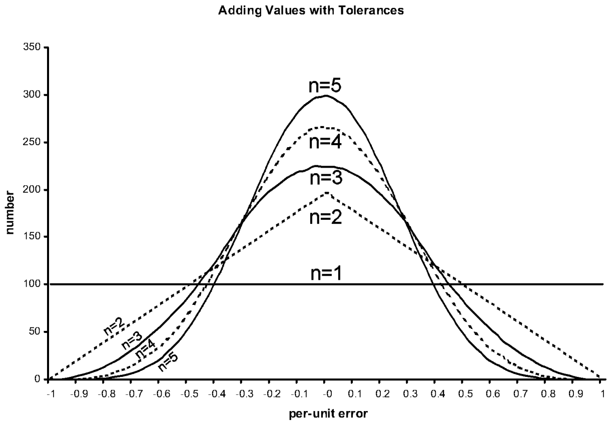
a) What is the worst case gain tolerance?
b) What is the statistical [RSS] gain tolerance?

RSS tolerancing is not safe for such a small group of components as this; a better scheme is needed. There are also situations when the tolerances are added along with the values, although these tolerancing situations are many time less frequent than those given above.

Suppose you have four different resistors in series, making up a larger value. Let's say they are 300K, 330K, 360K and 10K making up a 1M resistance. If they are all 1% resistors, what is the worst tolerance, 4%? No, it is 1%. But now these components are different from each other, so they are likely to be statistically independent. Neglect the effect of the 10K because it is 30× smaller than the rest. The other three are roughly equal values so they have roughly equal effects. The distribution of the total resistance will be *closer* to the nominal than that of the individual resistances, but how much closer?

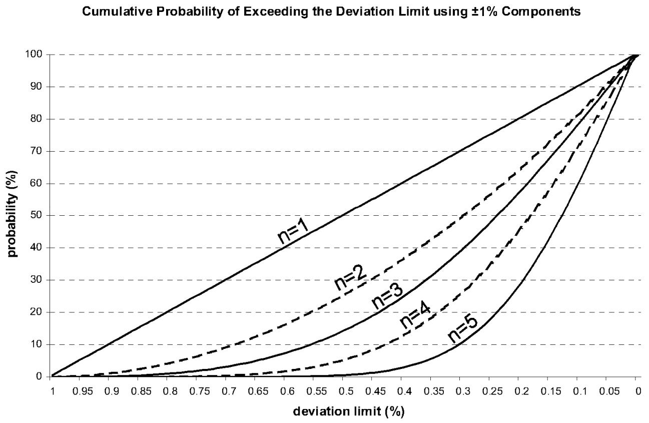
The *safe* assumption for individual resistors is that they have a flat probability distribution of resistance. In other words the resistor is equally likely to have any value within the tolerance band. It is unsafe to consider a component as having a Normal {Gaussian} distribution with ±2σ limits at the tolerance band edge. Sometimes the tolerance band is an allowance for systematic calibration uncertainty. Sometimes the tolerance band allows for differences between different manufacturing plants or different batches. Nothing in any manufacturer's data sheet ever allows you to assume that the specified tolerance has any specific probability density function.

FIGURE 3.4A:



n=1 is the flat distribution assumed for the individual part. As more of these are added together, you get something that looks more and more like a Gaussian distribution. This is the Central Limit Theorem in operation. There is definitely an improvement in the effective overall tolerance, but it is difficult to read any sort of value from the graph.

FIGURE 3.4B:



The cumulative probability chart shows that if you have 5 roughly equal elements averaged together, and each has a flat tolerance distribution of ±1%, then there is a 5% chance that the sum will exceed ±0.5% limits. Had you calculated a limit using the statistical method, the resulting $\frac{\pm 1\%}{\sqrt{5}} = \pm 0.45\%$ value would have been less safe.

Given that only the higher confidence levels are of interest, a table of values is useful. This table is for equal additive errors and shows how much of the error remains at the confidence interval shown. The numbers are ***per-unit*** amounts of the error remaining. With two components, each with a ±1% tolerance, you can only say that they are better than 0.900×(±1.0%) = ±0.9% overall, if you want to achieve a 99% confidence level.

confidence	n = 1	n = 2	n = 5	n = 10	n = 100
95%	0.975	0.776	0.502	0.356	0.113
99%	0.995	0.900	0.639	0.461	0.148
99.9%	0.9995	0.968	0.773	0.576	0.189

The results depend on the rules used to set up the simulation in the first place. If, instead of a flat probability density function for the initial values, a Gaussian distribution is used, a very different answer is obtained. The effect is to multiply the actual value of n by perhaps 2 or 3. The problem is that you are then trying to be more specific about the *unknown* and *unmeasured* distribution of the components.

Another way you could look at this problem is to use the Student's *t*-distribution tables. These take into account the low number of components in terms of the number of *degrees of freedom*. Effectively you need tables for the number of components you are using and the confidence level you are working to.

One problem you get into is that of *selection*. If you buy ±5% components, but ±1% components of the same type are also available from the same manufacturer, you may get a batch where all the 'good' ones (better than ±1%) have been selected out. In this case your components may **all** be worse than ±1%.

The worst case tolerance is overly pessimistic and the statistical tolerance is overly optimistic. The actual tolerance used should be somewhere between these two limiting values, at a position depending on the number of components involved and whether or not the tolerance is dominated by just a few large-tolerance components.

In the summation process it must be recognised that the tolerances being used are not the component tolerances, but the resulting tolerance on the output variable. A component tolerance requires a *weighting factor* { *sensitivity factor* } to refer its change to that of the output variable. If a ±5% change of a particular resistor gives a ±2.5% change of the system gain, the weighting factor {sensitivity} is 0.5×. You only combine component tolerances after each one has been multiplied by its sensitivity factor.

To sum component tolerances, add the 4 biggest weighted tolerances together, then add the RSS sum of the rest.

The justifications for this *reasonable uncertainty* rule are:

- it gives a value somewhere between the worst case and statistical measures
- it automatically compensates for cases where there are dominant tolerances
- it handles cases where there are not enough components to produce a good statistical measure
- it gives useful answers

Be aware that this is not an industry-standard rule.

EX 3.4.2: 10 individual unrelated resistors set the gain of an amplifier. They all have ±1% tolerances and unity sensitivity factors.

a) What is the worst case gain error?
b) What is the statistical [RSS] gain error?
c) What is a *reasonable uncertainty* of the gain?

A separate rule is needed for the case where the weighted tolerances are added then divided by a new mean value. Scaling the table given previously offers the most representative answers.

***EX 3.4.3:** A power supply has a load consisting of at least 100 different items. The load consists of logic ICs, opamps, transistorised circuitry &c. You don't know much more about it than that, other than that the PCB on which the parts are mounted is fairly evenly hot. Make an *engineering estimate* at the expected variation of supply current due to component tolerance alone.

***EX 3.4.4:** Two designers are working on two separate high volume (>100 units per week) projects. Each design has at least 50 critical tolerance problems to solve. Designer *A* uses a 99% confidence interval and Designer *B* uses a 99.9% confidence interval. What is the average production first time pass rate of designs from these two designers, based solely on this statistical data?

This exercise should demonstrate that *most routine design work should be done at a 100% confidence level*. By this I mean using *additive* tolerancing. Only on a few key expensive areas can you afford to be playing with the possible yield loss that working with lower confidence levels brings about. Power supply loading, however, is one case where statistical tolerancing is essential to get a reasonable cost.

3.5 Monte Carlo Tolerancing

During WWII, nuclear physicists were trying to solve problems involving radioactive decay interactions between large groups of atoms. Although the problem was analytically soluble for a single atom, the complexity of the multi-atom situation was far too great. The solution they came up with was computation based on the known equations, but including a random element {part, factor}; the random element being necessary to account for the non-deterministic decay of radioactive atoms.

The name *Monte Carlo* simulation, derived from the home of the famous casino in Monaco, gradually became popular in the late 1940's to describe any simulation process involving an element of chance. The technique itself, however, had been used by statisticians since the beginning of the 20th century under the name "model sampling".

For our purposes, Monte Carlo analysis is a 'brute force' computational technique. Rather than think about which components affect which part of the spec, the computer changes all circuit values randomly and records the results.

Does this mean that you should abandon the previous section and let the computer do it all? **NO!** The tolerancing I have previously described is the correct way to handle linear systems where the tolerance effects of components can be algebraically investigated. It is also worthwhile doing this same work by computer analysis, to see which components have the most significant effects.

With Monte Carlo analysis the computer can show the designer what is going on. By storing not only the output results, but also the input variables, it is possible to see the effects of individual component variations. It is up to the designer to utilise this tool in the appropriate manner. Simulating a whole system to investigate the gain variation of a single ×10 stage is remarkably inefficient. What is useful, however, is to do say 30 Monte Carlo runs of the whole system to ensure that there are no unexpected surprises lurking in the design.

Monte Carlo simulation is really needed in the design of circuits where the component values interact in a complex manner. This would not generally apply to DC operating point analysis. It is much more likely to occur on pulse response simulations and bandpass frequency response simulations, particularly where there is poor buffering between stages.

Again this is not something you would do for single-***pole*** or two-pole circuits. They are easily calculated. But more than two poles and the computer will suddenly become your best friend. Set up the circuit topology on a SPICE optimiser and let it run overnight, if necessary, to see if it can find a better solution. This is really quite efficient in terms of design time, as it does not cost much to leave a desktop computer running overnight.

A narrow band filter design typifies a useful application. This design could be for an intermediate frequency (***IF***) filter with a very sharp cut-off or it could be a notch filter to remove ***mains*** frequency noise. In either case, a Monte Carlo simulation can graphically tell you if, due to component tolerances, the design will continue to meet its spec. Such analysis would be impossible by hand, and laborious if you were to simulate the tolerance of each component one at a time. And don't forget that you would also need to consider the *interaction* of the component tolerances, which would be essentially impossible except by the Monte Carlo method.

It is all very well manually 'tuning up' a *notch filter* to 90 dB attenuation at the centre frequency, but when a capacitor in the filter drifts by 1% with time, how deep will the notch be then? In circuits of any complexity in this respect, computer simulation is the answer. If you look at the circuit in terms of reactive elements {inductors and capacitors}, then any more than four joined together, with or without resistors, is mathematically intense.

You can't say that increasing L1 will always lower the response at some frequency, so you are virtually forced to simulate the sub-circuit. On a large circuit consisting of several separable blocks, break the whole circuit down and simulate each block extensively. Only when each block is individually proved should you then simulate the whole system. However, don't neglect the final total simulation because there may be unexpected interaction between the blocks due to input/output impedance issues, for example.

3.6 Experimental Design Verification

Design verification is the last part of any design and yet it is right near the front of the book. It is a vital part of the design process and without this step the design cannot be considered complete.

Having done all the mathematics and measurements and so forth, a completed system has to be tested. Ideally several completed systems, but this is a question of economics and the number of units that will finally be produced. One or more systems have to be physically tried to prove that they meet the spec.

Ideally this evaluation would be done by a separate person (or team) from those doing the design. This is the ideal, because this nasty spiteful person (or team) has but one aim in life: to prove that your design is worthless, unfit for production and incomplete! It is difficult, but not impossible, to take this view on ones own design.

Suppose that the unit has to operate to 50°C. Do you test it to 50°C? NO! That would be madness. This one prototype does not contain the entire production spread of tolerances and yet all finished units have to pass the test. There are two possible schemes for this. One method is to test the unit to the spec limit and add a bit more as a safety margin; 5°C to 10°C would be a minimal margin. The second, and the better of the two methods, is to test the unit until it fails.

Now let me just clarify that a bit for you; don't necessarily destroy the equipment to show that it has failed. If a unit with a 50°C spec survives to 70°C, no testing beyond that point is necessary.

For an analog system you have *parametric failure* as a key concern. Perhaps the gain changes with temperature to the point where it is getting close to the spec limit. This would be very evident from the (analog) results taken. But with a system that contains digital devices, it is much more difficult to establish the margins. Unless you probe every interface and see how the timings change with temperature, you have no idea how close you are to any 'fatal' limit. This is the point of the 'catastrophic' failure testing. Test it until it 'falls over' and then, if it is close to the limit, or if time is available, find out what caused it to fall over. Whether or not you now do something to fix this failure is up to you; it is related to the nature of the failure, the cost to fix it, and how close the failure was to the spec limit.

Now remember that I am talking about a final system check. You must previously have checked the power-on and power-off cycles, basic timings, power supply currents; just the usual, everyday functions of the system. But as a final check you deliberately overstress the product to see which is the weakest point in the design. That it passes this final test is *not* a complete test of the design; it is *necessary but not sufficient*. Don't substitute this final test for all the intermediate testing. All that other design work still has to be done, but this last phase is a final system-level check.

On a project with many people involved, each should have individually checked-out their own part of the design. The system integration and verification stage checks out the interconnection of these parts. Was the spec of the individual parts adequate? This testing will highlight the problems.

Now remember that when you are going to stress-test a design, look for a worst case stress; if supply voltage causes significant changes use the worst supply voltage. Partially block the fan holes or push the equipment into the corner of a room so that the air flow is restricted. A *Taguchi* series of experiments is a good way of quickly establishing if there is a combination of factors that is going to give the design a problem.

I am not saying that you have to make equipment with a margin of 10°C on its operating temperature. I just want you to know what the limiting factor in the design is. And remember that what I am talking about is *any* spec, not just the ambient temperature. What happens if the mains supply voltage drops 2 V outside the permitted operating range; does the equipment 'drop out' gracefully, or does it catch fire and destroy itself? Catching fire and self-destructing is not going to be a useful mode for any marginally out of range input parameter!

You should always be thinking of 'what ifs?' when designing. For example, with thermally protected voltage regulators, if the power rating is exceeded the device may just shutdown. Thermal shutdown can be very unpleasant if devices on its output are running between multiple rails. The regulator is safe, but other bits may get destroyed as a result of the regulator shutting down. This is the sort of problem that is usually fixed by adding diodes between the power rails. Ordinarily the diodes are reverse biassed, but if one rail shuts down then these diodes conduct and stop other components getting heavily reverse biassed (and therefore destroyed). This sort of protection circuitry is seldom mentioned in circuit collections and application notes. Professional engineers need to include such parts in their designs right from the beginning, however, because faulty units returned from customers are very damaging to ones career prospects!

3.7 The Meaning of Zero

Zero is a definite integer quantity, but should not be applied to ‘floating point’ values.

There is no non-integer effect, quantity or measure which is zero.

It is unsound to describe an effect as any of the following without further quantification:

☹ negligible
☹ unimportant
☹ slight
☹ tiny
☹ minor
☹ irrelevant
☹ less than the noise
☹ too low to measure
☹ zero
☹ not relevant
☹ for all practical purposes it has no effect
☹ none

It is necessary to state just how low this *negligible factor* really is, or how low this *zero* really is. If somebody is measuring resistance with a hand-held multimeter, then 100 μΩ might well be negligible *for them.* If somebody else is measuring a 1 Ω resistor with great precision then this same 100 ppm error may well be completely unacceptable. The answer is to say that the effect is negligible by all means, but then append some *quantifying* statement to show the resolution of the measurement by which it was found to be negligible.

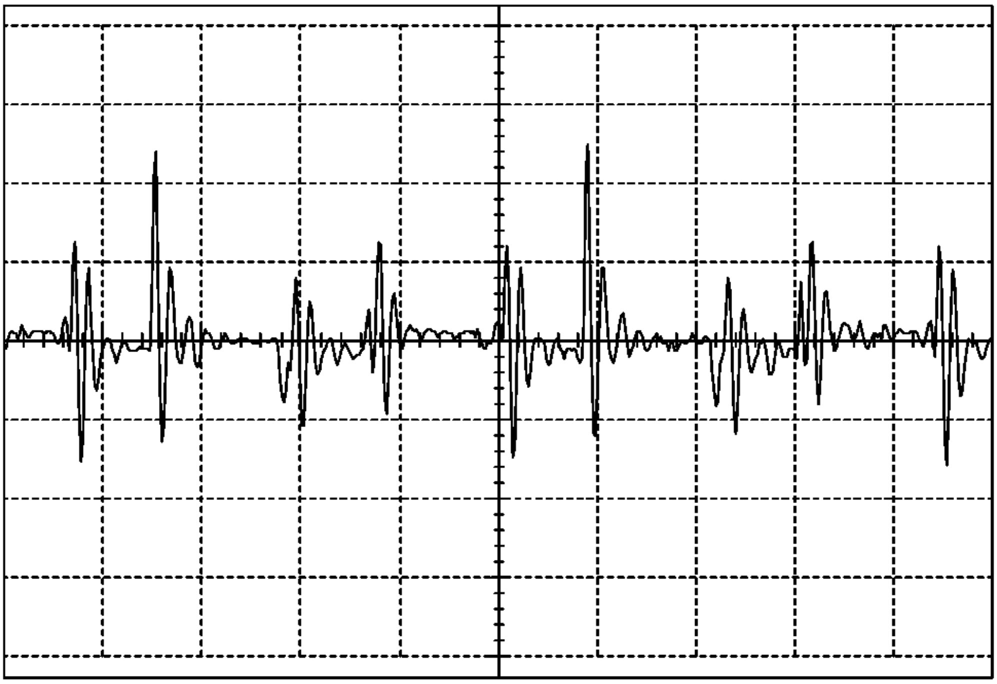
(As a side note, I originally had “immeasurable” on the above list, using the definition “that cannot be measured”. However the full dictionary definition of the word has a leaning towards the quantity being too large and boundless to measure. Hence immeasurable was removed from the list.)

Zero-effect or no-effect is an impossible condition for analog variables. It is always better to specify an effect as being less than some particular value. This type of statement is far more useful to an engineer.

Consider this real life example. A switched-mode power supply was producing a nasty spiky power rail which had an amplitude of about 46 mV ptp. On viewing this waveform with a scope the waveform had a dominant repetition period of about 45 μs. On the scope the mains frequency ripple seemed to be negligible, *in comparison with the switching frequency noise at around 22 kHz*. The thing is that the analog circuitry had decoupling capacitors which could heavily attenuate the high frequency switcher noise.

This noise, whilst having a repetition period of 22 kHz was in reality primarily at some much higher frequency (see actual measured waveform below).

FIGURE 3.7A:



10 mV/div
10 µs/div

The net result was that the high speed noise was effectively eliminated, but the mains frequency noise was unaffected and turned out to be the dominant power supply related noise source!

There is no such thing as a perfectly linear analog response

I get very suspicious when I see a perfectly straight line as a *measured* response. I want to see the *actual* data points as well as the ***least squared regression*** line that is put through them. This then shows the amount of noise on the measurements and gives one confidence that the line actually has some meaning. You can *always* draw a straight regression line for any number of points on a graph; it is just that the regression line will be virtually meaningless if the actual data points are scattered wildly either side of it.

CH4: the resistor

4.1 Resistor Types

A *resistor* is an electrical component having *resistance* as its dominant electrical attribute. Resistance is in turn defined by *Ohm's Law* of 1827: [1] $V = I \times R$. Components which follow this law are linear or *ohmic*; components that don't follow this law are non-linear or non-*ohmic*, the term non-linear being preferable.

In materials science and semiconductor studies, the term 'non-ohmic' is used to describe rectifying junctions. A non-rectifying connection to a semiconductor would be called an *ohmic contact*, but such a contact would not necessarily be entirely linear.

Is an ordinary inexpensive resistor linear? Well, that depends on how closely you measure it! Remember that there is no such thing as "zero" non-linearity. Even cables used for RF transmitters have their linearity checked and specified in terms of *passive intermodulation distortion* (***PIM***).

There are three fundamental types of resistor:

- film
- wire-wound
- bulk

This list covers all types of resistor that exist today, that have existed in the past, and that will exist in the future. Let's look at these in reverse order of importance, preventing the less relevant types from cluttering up your mind.

I have grouped all bulk resistance elements such as gaseous, liquid, and solid elements into one class. Thus conductive concrete, as proposed for de-icing of bridges, would come into this class. Also included would be "water resistors". For high surge-withstand capability (kJ pulses at kV levels) it can be convenient to make a resistor out of a column of water containing a small quantity of a water-soluble salt such as copper sulphate. Such a resistor is easy to adjust and experiment with. The large surface area of a column of water, perhaps 1 m long and several centimetres in diameter, makes it easy to dissipate the pulse energy. Furthermore, the column is a continuous system, having no region of excess electric field stress.

Carbon-composition resistors, moulded from a mixture of carbon and insulating binder, are virtually obsolete nowadays, but used to be the standard low-cost resistor prior to 1960. They were largely replaced by carbon film resistors circa 1980. They had horrible tolerances (±10% was common), poor long term drift characteristics (worse than 5%), large TCs (>1000 ppm/°C), they were electrically noisy and had poor behaviour above a few tens of megahertz. Their only virtue was that they had good pulse-withstand capability because of their solid construction. For example a ¼ W 1K resistor could be specified at a 10 kV pulse level.[2] This is not something you can do with an ordinary ¼ W film resistor. Nowadays you should use a high-pulse-withstand film resistor, rather than trying to get a carbon composition type.

[1] 'Georg Simon Ohm', CDROM edn (Encyclopaedia Britannica, 2000).

[2] SEI type RC.

Wire-wound resistors are often used for high power applications [>10 W], but they tend to be expensive. They are rapidly being replaced by “power oxide” film resistors, which are significantly cheaper. As their specialist application is high power, you will often find wire-wound resistors in heavy metal or ceramic bodies, designed to be clipped or screwed directly to heatsinks.

Since wire-wound resistors are coils of wire, they have relatively large (usually unspecified) inductances and are therefore not well suited to even moderate AC frequencies, above say 30 kHz. ***Bifilar winding***, a low-inductance winding technique, can reduce this effect. However, since the winding is more complicated, it is not generally done on ordinary power-resistors.

At the other end of the power usage range you find calibration laboratory resistors. If designed for any sort of power dissipation (say more than a few hundred milliwatts), they are generally in large oil filled cases to minimise the self-heating temperature rise. Some types are designed for immersion in a temperature controlled oil bath for the same reason.

These calibration-standard resistors cost so much anyway (say >$600), that the additional cost of bifilar winding is often acceptable. The special low TC wire is usually wound without a bobbin so that the wire is not under mechanical stress. This open construction also allows the oil to circulate freely around the wire. If the wire is wrapped around some sort of frame, then it is necessary to do so without putting the wire under tension. A wire under tension will stretch over time, changing its resistance; this is clearly not good for a resistance standard.

Notice that the oil will absorb a small amount of water and the resistance may then be affected by humidity. The humidity problem is overcome using the *Thomas type* standard resistor construction, where the resistive element is hermetically sealed in an inert-gas atmosphere.

Bifilar winding is only important for low-value resistors, however. For resistors below say 10 Ω the inductive reactance can be significant. Above 1 kΩ it is difficult for the inductive reactance to be a large fraction of the impedance. It is much more likely that for these higher value resistors, the shunt capacitance will cause the dominant frequency dependant change in impedance. Bifilar windings run the *go* and *return* conductors right next to each other to minimise inductance. This construction is therefore very poor on self-capacitance. There are several ingenious winding techniques that have been used to minimise reactance in wire-wound resistors,[3] but these are of less importance nowadays since thin-film resistors give cheaper and more accurate replacements.

4.2 Film Resistors

Film resistors are the most common and the cheapest available resistors. Film resistors are also made for ultra-precision applications, whereupon the cost may have increased a thousand fold. They come in two common types:

- planar
- tubular

[3] F.E. Terman, 'Nonreactive Wire-Wound Resistors', in *Radio Engineer's Handbook*, 1st edn (New York: McGraw-Hill, 1943; repr. London, 1950), pp. 43-44.

Both types can then either have wire leads attached or be surface mounted. The films can be placed on arbitrarily shaped surfaces, but tubular and planar films are the norm.

The films are deposited in two distinct ways. The first is known as *thick-film* and is a relatively crude process of screen printing conductive ink onto a surface. It is cheap, and therefore commonly done. The second method is known as *thin-film* and is done by vacuum deposition. It is more expensive, but produces a more stable resistance (say 100× more stable).

Neither method produces resistors with adequate absolute accuracy, however. Untrimmed the resistors could be as much as 30% low in value, the trimming process always increasing the value. This trimming is done by making the current path through the resistive element longer. For the tubular bodied resistors, a spiral is cut with a laser until the desired resistance has been achieved. For the planar resistor, a cut is made at right angles to the current path (*plunge cut*) for a coarse trim and then, as the correct value is approached, the laser cut may be moved parallel to the current path for a finer trim (*L-cut*).

The most important thing to realise about these trimming techniques is that they create great localised voltage stress regions in the film and therefore make the resistor less stable with time. Companies which manufacturer thick or thin-film resistors, and resistor networks, have their own proprietary rules for trimming them to achieve optimum stability. If the laser path is longer, the resistor costs more to trim and the laser can overheat the substrate, causing greater micro-fractures in the film.

The laser beam is not continuous, as you might have expected, but is actually a series of short high-energy pulses. The exact beam widths, energies and pulse durations are factors which manufacturers are constantly changing in an effort to improve performance. To give you some sort of feel for the situation, consider a spot diameter of 50 μm, 1 mJ per pulse, and a repetition rate in the kHz region. Thus a laser cut, also known as a *kerf*, is made by overlapping a series of these spots.

The problem with the laser is that the beam is not of constant intensity across the cutting spot, tending to fall away in intensity towards the edges. Thus in order to vaporise the material in the middle of the spot, the material around the edge will be over-heated but not removed. The result is a series of micro-fractures around the kerf, adversely affecting the stability of the trimmed resistor. It is therefore important to use an intelligent trimming process which not only measures the instantaneous resistance value, but also predicts where it will drift to in the first few hours after trimming.[4]

Clearly for VHF purposes the use of surface mount resistors is preferable, there being no lead-wire inductance to take into account. However, the spiralling of the standard tubular (MELF) surface mount resistors introduces additional (and variable from batch to batch) inductance. Manufacturers of micro-MELF resistors have therefore developed the pulsed helical trim technique to minimise this effect.[5] Imagine cutting a helix in the tubular film with a laser, but leaving large gaps in the kerf, making each helical cut no more than 30° of arc. This technique gives quoted performance up to 10 GHz. Having said that, rectangular 0302 resistors are available with quoted performance up to 40 GHz.

All of this high frequency trimming detail is most applicable to resistors below

[4] R Dow and others, 'Reducing Post-Trim Drift of Thin-Film Resistors by Optimizing YAG Laser Output Characteristics', *IEEE Transactions on Components, Hybrids and Manufacturing Technology*, CHMT-1. 4 (Dec 1978), pp 392-397.

[5] BC components BEYSCHLAG MMU 0102 HF.

1000 Ω, as mentioned previously. For a resistor of 10 kΩ and above, it is not possible for the series inductance to contribute a significant change in impedance. The frequency dependence of the impedance will then be most affected by the parasitic shunt capacitance in the resistor. This capacitance is in turn affected by the shape and width of any trimming cut.

Actually the bulk composition types are not much good for >10 MHz operation either. The current flows through the whole body at DC, but with AC the current flow is progressively reduced to the outer surface of the resistive element. This ***skin effect*** therefore suggests that conductors of AC are best when they have a large cross-sectional circumference, especially where the skin depth is small compared to the cross-sectional dimensions.

The calculation of the effective resistance of a conductor based on the product of the skin depth, the resistivity, and the cross-sectional perimeter length of the conductor, is a poor approximation for non-circular conductors. The actual resistance for rectangular conductors can be as much as double that calculated on the basis of their cross-sectional perimeters.

Planar surface mount resistors come in the following main sizes:

Type	length	width	power rating
0603	0.060 inch	0.030 inch	62 mW
0805	0.080 inch	0.050 inch	125 mW
1206	0.120 inch	0.060 inch	250 mW

The first two digits in the code are the length (in the direction of the current) in thousandths of an inch, with the second two digits giving the width in the perpendicular direction. Generally the current flows along the length of the resistor, but it is possible to build the resistor the other way around, thereby reducing the inductance and increasing the current handling capability. Other sizes also exist, such as 0402 and 2010, but these are less common. It is even possible to buy 0201 resistors, but again only for very specialised use.

Wire-ended resistors come in 125 mW, 250 mW, 333 mW, 400 mW, 500 mW, 1 W, 2 W, and 5 W ratings, amongst others. Different manufacturers make virtually identical looking resistors which have remarkably different power ratings (2:1 ratio). This is simply a matter of the operating temperature of the resistive film and the resulting long term stability. If the materials of the resistor can withstand a higher operating temperature, then the resistor can have a higher power rating. As the technology of materials advances, you should expect the size of resistors to reduce for a given power rating and so they will inevitably run hotter.

4.3 Pulsed Power

Simplistic resistor models can withstand arbitrarily large peak powers, providing that the mean power is not exceeded. Real resistors cannot withstand arbitrary peak powers though, and will either fail (open or short circuit) or drift significantly from their nominal resistance values.

***EX 4.3.1:** A 68 Ω resistor has a rectangular voltage waveform applied across it, the mark/space ratio being 1/7. The cycle time is 16.66 ms. The peak voltage is 10 V and the minimum is 0V. Forget the tolerancing, drift, self-heating TC &c.

a) What is the peak power in the resistor?

b) What is the mean power dissipated in the resistor?
c) What is the RMS power dissipated in the resistor?
d) What is the RMS voltage across the resistor?

Pulsed power handing capabilities of resistors are not given by all manufacturers. You may actually have to specifically request the information from the particular manufacturer you are using. You can test a specific component to see if it fails, but this can only show that it is not suitable. If it passes the test, you still can't be sure. A resistor that has been more heavily trimmed will be more susceptible to failure; the one you tested might not have been trimmed at all!

You can 'play safe' and put in a resistor that can dissipate the peak power continuously, but if the ***duty cycle*** is 0.01, the resistor needs 100× the mean power rating! It is not very efficient design to use a 25 W resistor when a 0.25 W device would do the same job.

Fast pulse cycles are easier to handle than slow cycles since there is a thermal time-constant involved. What you should expect is that bigger, higher power components will have a longer time-constant and therefore they should be more able to withstand low frequency power cycles than smaller resistors.

The failure of the resistive element is simply due to it reaching too high a temperature at one or more points. Obviously you must take into account the ambient temperature and the mean power that is being dissipated in the resistor when calculating its peak power handling capability. If the mean power is higher, the peak handling capability is obviously reduced. Now you need to obtain or invent some numbers to work with, electronics being a quantitative subject.

For a rectangular waveform of duty cycle D and peak value $\hat{V}$:

$P_{mean} = D \cdot \dfrac{\hat{V}^2}{R}$ This can be alternatively expressed as $P_{mean} = D \cdot P_{peak}$ giving

$P_{peak} < \dfrac{P_{rated}}{D}$. This is a *necessary but not sufficient condition.*

The limiting conditions depend on the resistance value. Suppose that the maximum voltage rating (*limiting element voltage*) of that style of resistor is 500 V. This voltage across a 1 MΩ resistor is only 0.25 W, whereas 500 V across a 10 Ω resistor is 25 kW. When looking at the peak power rating of a resistor, the first thing to do is to calculate the limit imposed by the maximum voltage; exceeding the maximum voltage for more than a few microseconds is not a good idea.

Suppose that the resistance is sufficiently low that the power maximum is not due to the limiting element voltage. Further suppose that a standard 0.25 W resistor is given a 25 W peak power at a duty cycle of 0.001; mean power = 0.025 W. Will the resistor survive? Suppose I am going to apply the power for 10 seconds every 10,000 seconds. That is the duty cycle of 0.001 mentioned earlier. How many 250 mW resistors will withstand 25 W for 10 seconds? The answer is none!

At any given temperature there is a maximum steady power that can be dissipated in a resistor. This will heat the resistive element to its maximum temperature limit. If this maximum steady power is maintained on average, but with a pulsing waveform, the element temperature will cycle above and below the maximum temperature limit. Thus *the mean value of an acceptable pulsed power load must always be less than the* *acceptable steady state power*.

When a massive pulse power is applied to the resistor, the resistive element will start to heat up rapidly, the exponential rise initially appearing to be linear. Suppose that for a given steady applied mean power, the resistive element eventually settles to ΔT above ambient. Simplistically, on the overload, the element temperature will initially be linearly 'aiming' for $\frac{\Delta T}{D}$ at the same applied mean power. If the time of application is short, then the resistive element will not have time to heat up very much. This is related to a time constant for the heating process.

Now this 'time constant' is not a single-valued item. For the very fastest pulses there will be conduction in the resistive track itself. For longer pulses there will be conduction into the ceramic body of the resistor. For still longer pulses there will be conduction into the PCB via the leads for a wire ended device, or via the pads for a surface mount device. However, in order to get a workable rule I am just going to simplify this complexity and use a single $\frac{\Delta T}{t}$ rate.

This is easier with numbers. Suppose the resistor would heat up by 100°C at the maximum continuous power level. At full mean power and a duty cycle of 0.01 it would initially be aiming for 10,000°C. Let's suppose it would get there at the initial rate in 100 ms. That is a rate of 100,000°C/s. Allowing the element to get say 10°C above its mean level sets the maximum on-time as 100 μs. According to this simple idea I can apply 100× the rated power for 100 μs, 1000× the rated power for 10 μs and 1,000,000× the rated power for 10 ns. This is getting ridiculous. The maximum voltage rating will limit this 1,000,000× overload, but nevertheless the rate of growth seems too high.

When confronted with this situation, the correct thing to do is to select a resistor which has a pulse rating specified. These do exist from several manufacturers, quoted specifically for high pulse applications. That is the correct thing to do. Unfortunately, these high pulse resistors may not meet your requirements in terms of absolute accuracy, temperature coefficient, cost, impedance characteristics above 10 MHz, size, availability &c. Now what? You are back to engineering judgement and compromise. Is there a good chance that the component you want to use, although not ideally specified for the task, will actually work reliably in the application?

It is a necessary but not sufficient condition that it works on prototype units. Perhaps you can deliberately run it harder than it would have to work in its final application in order to give you more confidence. Additionally you can get data from components that *are* specified and try to make up a working spec for this particular resistor using this additional information.

Remember that the manufacturer of the high pulse-withstand resistors has had to characterise those as well; it is therefore not unreasonable for you to characterise a particular component for your application, provided only that the manufacturer does not subsequently change their process!

The formula given below is a kind of last chance to get some sort of a spec. Notice that the square root of the duty cycle has been used (somewhat arbitrarily) to limit the rate of rise of peak power as the pulse gets narrower.

$$\boxed{P_{peak} < \frac{P_{rated}}{\sqrt{0.0006+D}} \cdot \frac{1}{1+\left(\frac{f_K}{f_{rep}}\right)^{1/4}}}$$

This equation is only valid for $f_{rep} \geq f_K$.

f_K is a constant for the resistor, perhaps in the range of 0.1 Hz to 10 Hz.

f_{rep} is the repetition frequency of the pulse.

EX 4.3.2: A 250 mW resistor is pulsed with 10 W for 1 μs every 100 μs. Take f_K as 1 Hz. Is this pulsed power level safe according to the formula? Would you do it?

You may be called upon to design resistive elements for thick or thin-film resistors. If you want to make them as resilient as possible then minimise the *stress ratio*.

$$Stress\ Ratio = \frac{Peak\ Electric\ Field\ Strength}{Mean\ Active\ Field\ Strength}$$

The *mean active field strength* is the mean electric field strength for the area of the film that has an electric field strength greater than 5% of the peak electric field strength. This definition is necessary to prevent relatively unused portions of the film from affecting the ratio. The best possible stress ratio is 1. For an untrimmed pattern you should definitely aim for a stress ratio below 2.5 if you wish to maintain stability and achieve resilience in your resistor.[6] Furthermore P- and L- cuts are definitely to be avoided. A proposed method is the *lazy-J* cut,[7] this technique dramatically reducing the peak electric field stress in the resistor.

In a particular 0.5 W rated microwave attenuator design, the pulse handling was improved from 50 W for 1 ms to 50 W for 15 ms by subtle reshaping of the field distribution.[8]

4.4 Simple Design Exercises

Graduates who can't even light an LED are a liability. Here is your first design exercise. (If you are a practising designer then just read over the answer.)

EX 4.4.1: It is required to light up the "power ON" LED from the +12V power rail. The LED should be bright, so run it at 25 mA. At 25 mA this particular LED has a forward volt drop of 1.4 V. What resistor should be used? Consider *everything* relevant to the problem.

This task is a very common thing to have to do. It should not be a major exercise. You should have been able to complete the task in a few minutes on the back of an envelope. Actually in the real world it would have been harder. You might have had to select the LED and/or find a data sheet on it. If your company hadn't already been using that type of LED, you would have to take out a part number, write out a description, correct

[6] L.O. Green, 'Beautiful Resistors', in *Electronics World*, 107, no. 1779 (March 2001), pp. 194-197.

[7] L.O. Green, 'Reducing the Stress in Planar Resistors', in *Electronics World* (Dec 2003), pp. 12-16.

[8] N. Thomas, 'Improvements in or Relating to Attenuators', *UK Patent Application 2158999A*, *US 4672336* (UKPO, 1985: USPO, 1987).

somebody else's typing of the description, get samples, get quotations for the price and delivery, and just when you thought you were finished, a Senior Engineer asks if you have another manufacturer's part listed in case there is a supply problem with the first manufacturer!

This next exercise has completely different rules; if you have understood your basic theorems it will not be difficult. I am going to simplify the problem for you by letting you use arbitrary (non-preferred) resistor values and by neglecting tolerances, temperatures, power ratings &c. Just do it as an "academic exercise" to see if you can solve what is, after all, a very simple real world type of problem.

***EX 4.4.2:** A control voltage from a DAC has a range of ±5 V. Scale this ±5 V control range to swing from 0 mV to –100 mV in order to control a variable gain amplifier. "Noise-free" power rails of ±12 V are available.

Note the widely used (slang) terminology for attenuating a signal or control voltage is to "pot it down", meaning to use a resistive potential divider to reduce a signal amplitude.

4.5 Resistor Matching

In the section on combining tolerances you were told to always *weight* a component tolerance by a *sensitivity factor* for the circuit. This sensitivity factor must be evaluated either analytically, or by using a computer simulation. In this chapter the sensitivity factors are easy to evaluate analytically because of the simplicity of the circuits. The sensitivity factors can be evaluated either by *partial differentiation* or by substitution of a factor such as $(1+\delta)$ into the network equation, where δ represents the per-unit deviation of the value from its nominal value. For a resistor of value R in a voltage divider, for example, you could substitute $R \times (1+\delta)$ and evaluate what the effect was on the output voltage without using calculus.

@EX 4.5.1: Make a voltage divider out of a series-connected pair of resistors. Assume that the voltage source driving this divider has no resistance and is some exact value. Neglect all imperfections in the resistors other than their initial tolerance. Both resistors have a ±1% tolerance and they are statistically independent from each other. What is the worst case tolerance on the unloaded output of the voltage divider?

Once you have mastered that exercise you will have a very much better understanding of the subject of tolerancing. The question is always, how much effect does a specific component have on the final output? If it has very little effect then you don't have to specify an expensive component. If you can get the same functionality as an existing design, but can take 10% off the parts cost for little design effort, and no added test time, you will be popular at work.

At the beginning of this chapter I posed the question: Is a resistor *ohmic*? This could be answered in two ways:

➢ A good resistor follows Ohm's law reasonably well over a limited range of current.
➢ No resistor follows Ohm's Law if you measure it closely enough.

The first answer is arguably the most useful one, but the second answer is also correct.

The following factors affect the measured value of a resistor. The intention is that the most significant factors are nearer the top of the list, but for any particular design of resistor, and for any particular environment, it is never certain which will be the dominant factors.

☹ Temperature and Time
☹ The applied voltage or current
☹ Time since application of V or I
☹ Oxidation / Corrosion
☹ Mechanical Stress and Strain
☹ Humidity
☹ Recent changes in temperature
☹ Chemical decomposition
☹ Atmospheric pressure
☹ Vibration
☹ Gravitational Force

***EX 4.5.2:** A 10:1 voltage divider has been made from two individual resistors of 900K and 100K. Neglect the source impedance, load impedance, initial tolerance, long term drift, and non-linearity in the resistors. All environmental conditions are held constant. Suppose this divider produces an exact 10:1 ratio with a small voltage applied. How far off could this ratio be when 400 V is applied to the divider? Take the TC of the resistors as ±100 ppm/°C and the effective thermal resistance of each resistor as 200°C/W. The resistors should be considered as thermally isolated from each other for the purposes of this problem.

The numbers I have given above are quite realistic and it is a good example of a *self-heating error*. This effect will have a thermal time constant. There is another effect which I implicitly told you to ignore. This is known as *voltage coefficient of resistance*, often abbreviated to just voltage coefficient, and quantifies the non-linearity of the resistor.

The error due to a finite voltage coefficient is an instantaneous effect and is distinct from the self-heating effect. The voltage coefficient may be very small for accurate resistors; figures of 1 ppm/V are not unusual. Depending on the size, quality and area of the resistor, one effect can dominate the other. Vishay high voltage surface mount chip resistors type CRHV have voltage coefficients in the range 10 ppm/V to 25 ppm/V for resistors of 2 MΩ to 50 GΩ; this is thick-film technology. Vishay CNS471 precision voltage dividers on the other hand are specified as <0.002 ppm/V; this is thin-film technology.

In practical implementations of such an attenuator there are several things that can be done to reduce the linearity error to an acceptable level. If you make the attenuator out of 10 identical 100K resistors then they will all experience the same voltage and power. If they are all from the same batch then you would also expect that their TCs would be similar.

Formally you would call this *TC matching*, *TC tracking*, or *ratio TC*. A mathematician would say that the TCs were correlated. For a custom thick-film network you might state that the absolute TC of the resistors was 100 ppm/°C, but that their tracking TC was 10 ppm/°C. [The term *absolute* here means "not compared with anything else" and is used to distinguish between actual TC and tracking TC.]

If you were just making an attenuator from off-the-shelf identical components you might write down an estimate of the tracking TC. This is an engineering judgement. You cannot guarantee that the TCs will track, but you should have a reasonable expectation that they would.

In the worst case, the bottom resistor in the chain could have a TC of −100 ppm/°C and all the others could have TCs of +100 ppm/°C, but that is pretty unlikely. I would be fairly comfortable with using a tracking TC of ±20 ppm/°C, a 10× improvement. If you expected the ±100 ppm/°C TCs to track to better than ±5 ppm/°C then I would say that was pretty risky. Tracking is a matching effect, where you expect components made from the same materials at the same time to be reasonably well matched in terms of the nature of their response.

To guarantee the spec, if it is critical to the system performance, you would have to spend considerably more money by:

- buying a matched pair or set of resistors
- using low absolute TC resistors
- specifying your own thick or thin-film resistor network
- buying a standard thick or thin-film resistor network
- putting on an adjustment for TC and trimming after measuring the response.

Now you see why you might want to take a risk on the matching that you could get for 'free' using your engineering judgement. Using matched sets of loose resistors is an old fashioned way of going about this matching. Somebody has to physically match up the resistors and put them in a bag. Then somebody else has to take them out of the bag and place them in or on the board. For the ultimate in performance this sort of technique may be necessary. Otherwise it would only be acceptable if labour costs were low.

If you are using simple ratios like 2:1, 5:1 or 10:1 you will find that distributors may provide these as part of their standard range. If you need an obscure ratio then you will either have to have your own network made as a semi-custom design, or you will have to buy some precision resistors.

You can buy 0.1% 10 ppm surface mount resistors for say $0.80 each in small quantities. That's 80× the price of simple 1%, 100 ppm resistors. [I dropped the ± symbol and the /°C on the TC spec. It is a very sloppy way of describing a resistor, but it is also common practice.] Practically, I can make my attenuator with perhaps 3 or 4 off 1% resistors for say $0.04 (900K is not a preferred value. I would have to make that value using series-parallel combinations.) The 0.1% resistors I mentioned only go up to 100K. Above that you have to get them specially made or go to large wire-ended components. The cost is going to be well over $2.00. The question you have to answer for yourself is: are you willing to accept a TC of probably 20 ppm for \$0.04, or are you going to spend \$2.00 to guarantee the TC of 20 ppm?

For medical, military, and key spec points, you should definitely go for the guaranteed TC. For commercial applications on secondary spec points, you may consider that the cheaper solution is more appropriate.

4.6 Precision Resistance

It is quite common to have to trim a precision resistor to the exact value that you want. The question is: do you need to use a precision resistor to trim a precision resistor? Will using a 1% resistor in series or parallel with a 0.01% resistor ruin the accuracy of the system? This is back to the sensitivity issue mentioned previously.

***EX 4.6.1: A** 10:1 voltage divider is made from two individual 0.01% 5 ppm resistors. The top resistor in the chain is 9.000 kΩ. There are one hundred pieces of equipment incorporating this network on the shop floor and the quality manager has spotted that the overall system output is 1% low. Disaster; the customer will reject them! If these units aren't shipped today, nobody will get paid this month.

Some idiot from the shop floor has suggested that you put a 1 MΩ 5% 200 ppm resistor across the 9.000 kΩ, thereby trimming the overall system response back to nominal. The Boss asks you whether or not it will work. Before you can answer, the shop floor worker shouts out that he has already tried the modification and that it does in fact work.

Is it acceptable? Decide now, but be ready to defend your decision when the Chief Engineer returns from holiday.

This is another general class of problem that comes up time and time again. You will note that the effect on the TC, the absolute value, and the long-term drift all follow the same formula. When the trim amount is as small as 1% you don't have to try very hard to work out the answer. Just add 1% of the imperfections in the trimming component onto the main component values. Thus if the TC of the trimming resistor is 200 ppm the combined resistance will have a TC 2 ppm worse than the precision resistor's TC. There is no statistical weighting, you just assume that they add in the same direction to give the worst case.

You could have worked out that problem very accurately, but there wouldn't have been much point. The original gain TC was 9 ppm and the new TC is roughly 11 ppm. That isn't ideal, but may be acceptable. However, the initial tolerance of the 5% resistor is going to contribute 0.05% to the combined resistor. Instead of a 0.02% ratio you now have 0.07%. That looks unsafe. If somebody has gone to the trouble of putting 0.01% resistors in the circuit then making this more than 3× worse looks unsound. The 5% resistor will cause the combined resistor to drift more with time as well. A better quality trim component is needed; 0.1% 50 ppm would be good, 1% 100 ppm might be acceptable. You can formalise this method by writing the per-unit trim amount as T:

Total TC = (1−T) · Main TC + T · Trim TC
Total tolerance = (1−T) · Main tolerance + T · Trim tolerance
Total long term drift = (1−T) · Main drift + T · Trim drift

Because T was 0.01 in the example, it was easier to just consider (1−T) ≈ 1. This is not such a good approximation as $\mathcal{T}$ gets above 0.1 and you will definitely need to consider using the above formulae.

There are very few truly random events. When you measure a resistor from day to day and its value seems ‘random’ to some degree, even though the measurement system is relatively noise free, it may well be that you are just not taking into account significant factors that are changing the value. These factors could be contact resistance, humidity, ambient lighting and atmospheric pressure, most of which are neither measured nor taken into account. One good thing that can be said about atmospheric pressure is that at a fixed location, the atmospheric pressure change is never greater than a few percent.

An effect which will appear time and time again is ***thermal hysteresis***, also known as *thermal retrace*. If you heat a component up and then cool it down to the same temperature, the value is unlikely to be exactly the same. This effect occurs on all components to wildly varying degrees. It is one of those effects that you can essentially ignore on 5% components, but is usually significant on 0.001% (10 ppm) components. This thermal hysteresis effect also applies to capacitors, *Weston Standard Cells*, zener voltage references, and so on. You should expect to find this sort of effect on any system, and should be happy if the value is demonstrably small. It is only recently that voltage references are having this figure quoted by manufacturers. The effect has always been present, it has just recently become commercially expedient to mention it.

Thick and thin-film resistors have special rules for getting good results. If you are designing Nichrome elements on an IC, or a thick-film resistor on a ceramic substrate, the design considerations are similar. The thickness of the film is process-dependant and is not something that can be improved by the layout. This tolerance is usually at least ±15%. Planar film resistors are specified in terms of Ω/□, read as ‘ohms per square’. The resistance between the ends of a rectangular box element is given by:

$$\boxed{R = \frac{\rho \cdot L}{W \cdot T} = \frac{\rho}{T} \times \frac{L}{W}}$$

The Ω/□ figure is simply the ρ/T term, the resistivity divided by the thickness. All you get to adjust is the length/width ratio, *L/W*.

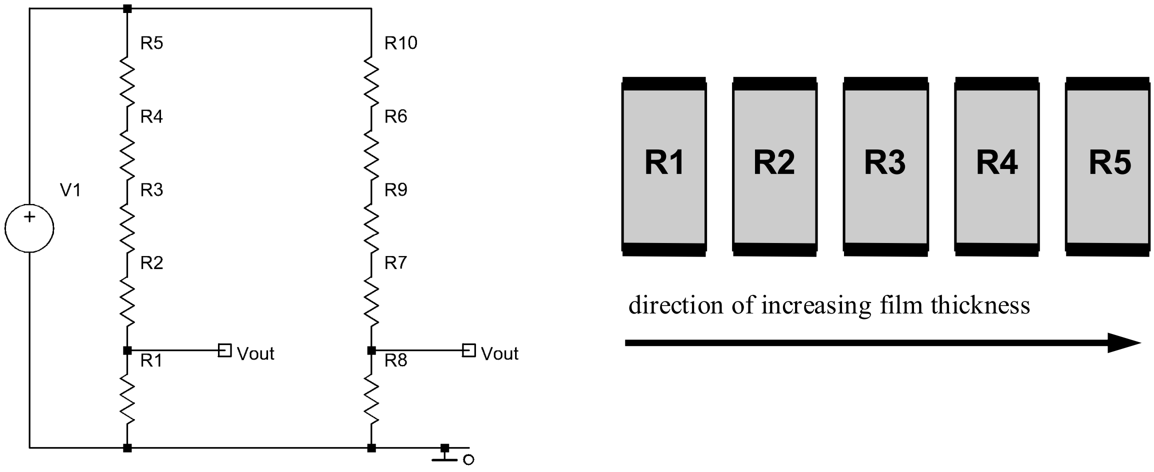
Some ‘learned texts’ now discard the “square” part of Ω/□, giving the surface resistivity in ohms. However, although “square” has no dimensions, and is clearly not an SI unit, it is widely used and understood in industry. Everyone will understand Ω/□, whereas surface resistivity in ohms is certain to cause confusion.

When a film is deposited, it is supposed to be of uniform thickness. If the mask that defines the resistor has equal apertures in it, the resistors will all be nominally equal. If there is any systematic thickness distribution created by the deposition process, that fact can be exploited to get better performance for little extra cost. Note carefully that I am *not* suggesting that it is good to have systematic errors in the process, but if such errors do exist they can be minimised by careful design. The primary application is for resistors that are not trimmed, diffused resistors in an IC being a good example.

To get the best matching between resistors on the same substrate it is essential that they are close together, have the same size, and the same orientation. Being ‘close together' does not mean <1 mm apart, it is a relative requirement. Remember that the whole die of an integrated circuit might only be 1 mm on a side and you would still want to place matched resistors as close together as possible.

Resistors of different sizes will not match well because process variations such as longer diffusion time or over-etching will make the L/W ratio change differently according to both the size and the aspect ratio of the resistors. Ask the manufacturer to guide you on the quantitative aspects of their process variations.

FIGURE 4.6A:



EX 4.6.2: Is there any difference in the accuracy of the two 5:1 attenuators shown above? Hint: Assume that the thickness gradient shown on the diagram makes R5 5% lower in resistance than R1.

It is quite possible that the systematic error in the second case is swamped by random factors in the resistor values. Nevertheless, the resistor matching has been improved without adding anything to the part cost. This sort of procedure is widely used in integrated circuits. In the input stages of opamps, for example, the input stages can be connected in such a way that gradients of temperature, doping density, oxide thickness &c have a greatly minimised effect on the overall result. This technique is known as *common centroid* design.

The connection of resistors in this way is most relevant to untrimmed resistive elements. If the elements are laser trimmed then it is wise not to have too many of them. Each resistor needs to be actively trimmed {trimmed whilst being measured} and this can add a significant amount to the cost.

The long-term drift of resistance value is dependant on how hot the resistor gets. If the resistor is run at its full rated power for thousands of hours the spec is called *load life stability*. If the resistor is run at <1/100th of its rated power, you can apply the *shelf life stability* spec. However, if the resistor has been heated by a wave-soldering machine (for leaded components) or an infra-red reflow machine (for surface mount components) the resistor may drift excessively in the first hours after this thermal stress.

This additional drift is not specified and can only be evaluated experimentally. For this reason it is wise to let circuits 'settle down' for a day after the soldering operation before final calibration adjustments. This is not ordinarily a problem for instrumentation which goes through initial board testing, gets assembled, then is burnt in for a 24-48 hours to eliminate *infant mortality*. By the time the loaded PCBs get to final test and calibration they can easily be more than 'one week old'.

On a high-speed, high-volume production line that builds, tests and calibrates product in less than one day, it would be wise to ensure that the calibration of the finished boards was adequately stable compared to the end-user spec. Such verification could be done by making measurements spaced one day apart.

Resistors have shunt capacitance related to their body size. Thus making a resistance out of several resistors gives a better high frequency response, according to some authors at least.

***EX 4.6.3:** A 1206 surface-mount resistor has 0.05 pF of shunt capacitance. Compare the frequency response of a single 10 MΩ resistor of this type to a chain of five 2 MΩ resistors *when used in a real application*.

4.7 Johnson Noise

Johnson noise[9] is the random, but statistically predictable motion of charge carriers within a resistive material. Think of the charge carriers moving about randomly, like gas molecules within a container. The random nature of this system allows electrons to group together at the ends of the conductors to some extent. There will effectively be a random voltage appearing across the resistive element.

Johnson's original paper is an experimental investigation of noise in all sorts of resistive substances. The theoretical explanation was published at the same time by Nyquist.[10] Some authors therefore prefer the name 'Nyquist noise'; others prefer *thermal noise*, the noise power being proportional to absolute temperature.

In Nyquist's paper the theoretical basis for the noise is thermodynamic equilibrium between two conductors at the same temperature. The mean power flow in each direction is required to be equal in order to maintain the equilibrium. A further consideration is that the mean power flow in any small bandwidth must also be equal, since a band-pass filter could be inserted into the connection without affecting the equilibrium.

The complete formula for the *available noise power* {the maximum noise power that can be extracted from a source} is:

$$\boxed{P_N = \frac{hf \cdot \Delta f}{\exp\left(\frac{hf}{kT}\right) - 1}}$$

h is the Planck constant, $6.6 \times 10^{-34}\ \mathrm{J \cdot s}$

k is the Boltzmann constant, $1.38 \times 10^{-23}\ \mathrm{J/°K}$.

For ordinary temperatures (≥150°K) and frequencies (≤100 GHz), $\frac{hf}{kT} < 0.032$

The power series approximation $\exp\left(\frac{hf}{kT}\right) \approx 1 + \frac{hf}{kT}$ is then accurate to better than 0.1%,

[9] J.B. Johnson, 'Thermal Agitation of Electricity in Conductors', in *Physical Review*, 32 (July 1928), pp. 97-109.

[10] H Nyquist, 'Thermal Agitation of Electric Charge in Conductors', in *Physical Review*, 32 (July 1928), pp. 110-113.

giving: $P_N \approx \dfrac{hf \cdot \Delta f}{1 + \dfrac{hf}{kT} - 1}$ $\boxed{\therefore P_N = kT \cdot \Delta f}$ The power spectral density is $\dfrac{P_N}{\Delta f} = kT$

At room temperature *kT* is 4.1×10^{-21} W/Hz, which translates to −174 dBm/Hz. This noise is both real and demonstrable, provided the right equipment is available. If a low-noise amplifier has its input connected to a screened termination resistor and the output is fed into an RMS responding voltmeter (or a power meter), the noise will be seen to reduce when the input resistor is cooled. Liquid nitrogen is easy to work with and can quickly cool the resistor down to 77°K, giving a large ratio of hot to cold input noise power (290 / 77 = 3.8). The ratio of output hot to cold noise powers is called the *Y-factor* and can be used to establish the broadband ***noise figure*** of the amplifier.

Using the Thévenin equivalent circuit, the maximum available noise power from a resistor is equivalent to a voltage source in series with the resistor:

$$\boxed{V_N = \sqrt{4 \cdot kT \cdot R \cdot \Delta f}}$$

V_N is the RMS noise voltage.

T is the thermodynamic temperature in degrees Kelvin.

k is the Boltzmann constant, 1.38×10^{-23} J/°K.

R is the resistance in ohms.

Δf is the frequency measurement band, assuming a ***brickwall filter***.

The formula gives a fundamental uncertainty on a voltage measurement in a given bandwidth. This is one reason why wideband instruments, such as scopes and spectrum analysers, are less accurate than their DC/LF equivalents. In order to get a reduced uncertainty it is necessary to use a lower measurement bandwidth. This can be done by analog means, for example by adding a large capacitor to form a low-pass filter, or by digital means, such as averaging.

4.8 Excess Noise

All resistances have Johnson noise, regardless of the conducting element used. However, real-world resistors have additional noise when current flows through them, *excess noise*. The first detailed paper on this subject was published in 1934.[11] Because excess noise is proportional to current through the resistor, it is also known as *current noise*. Whilst Johnson noise has a flat frequency distribution, excess noise has a 1/f characteristic, also known as *flicker noise*.

To summarise, current flow through a resistor creates noise, the bigger the current the bigger the noise. This additional noise is known as *excess noise*, *current noise*, 1/f *noise* or *flicker noise*.

The 1/f characteristic gives equal noise power per decade of bandwidth. Thus the noise power in the decade from 1 Hz to 10 Hz is equal to the noise power from 0.1 Hz to 1 Hz. Flicker noise is clearly of great significance to sensitive measurements where the bandwidth is reduced to minimise the noise. For flat spectrum noise, if you reduce the bandwidth by a factor of 100 you reduce the noise power by a factor of ×100 and the

[11] R.H. Campbell, and R.A. Chipman, 'Noise from Current-Carrying Resistors 20 to 500 Kc', in *Proceedings of the IRE*, 37 (Aug 1949), pp. 938-942.

noise voltage by $\sqrt{100} = \times 10$. If 1/f noise is dominant, then reducing the response from the 0.1 Hz-1000 Hz band down to the 0.1 Hz-10 Hz band will only reduce the noise power by a factor of 2, and the noise voltage by a factor of $\sqrt{2}$.

As excess noise increases with bias voltage, it is usually expressed as $\mu V / V$ on data sheets (although $\mu V / V / \sqrt{decade}$ would be more appropriate). Only good manufacturers quote excess noise, and their noise is therefore likely to be lower than the (un-quoted) noise of their competitors.

Excess noise is due to the irregularity of the current path through the resistor. An irregular path gives an irregular current and therefore more excess noise. Better processing, resulting in smoother surfaces on resistive elements, therefore gives less excess noise. Metallic films have considerably less excess noise than films made with conductive particles dispersed in insulating material. Thus higher valued resistors are noisier than lower valued resistors. Plotting data from one manufacturer [12] gives the excess noise increasing roughly as the square root of the resistance.

For the same type of resistive film, smaller resistors will be noisier than larger resistors, the noise power being inversely proportional to the active area of the resistive film.

A technique of measuring the noise was developed around 1960,[13] this method being used as the basis of the IEC and MIL-STD 202 (method 308) tests. The IEC standard circuit for measuring current noise [14] requires two resistors which are “current-noise free”; good quality wire-wound resistors satisfying this requirement.

For ordinary thick-film surface mount resistors, the generic specs are:

resistance	Noise	Noise Index
R ≤ 1 kΩ	≤ 1 $\mu V / V$	≤ 0 dB
1 kΩ < R ≤ 10 kΩ	≤ 3 $\mu V / V$	≤ 9.5 dB
10 kΩ < R ≤ 100 kΩ	≤ 6 $\mu V / V$	≤ 15.6 dB
100 kΩ < R ≤ 1 MΩ	≤ 10 $\mu V / V$	≤ 20 dB

Another way of expressing the excess noise of a resistor is to give the ***noise index*** (also known as the current-noise index).

$$Noise\ \ Index = 20 \times \log_{10}\left(\frac{RMS\ \ noise\ \ voltage\ \ in\ \ one\ \ decade\ \ of\ \ bandwidth\ \ [\mu V]}{DC\, bias\ voltage\ \ [V]} \right)$$

Because of the use of μV on the top and V on the bottom, an excess noise value of $1 \mu V / V / \sqrt{decade}$ corresponds to a noise index of 0 dB. A noise index of –54 dB at

[12] Phycomp RC02 1206 resistor.

[13] G.T. Conrad, N. Newman, and A.P. Stansbury, 'A Recommended Standard Resistor-Noise Test System', in *Institute of Radio Engineers: Transactions on Component Parts*, CP-7, no. 3 (Sept 1960), pp. 71-88.

[14] 'Method of Measurement of Current Noise Generated in Fixed Resistors', IEC 60195 (International Electrotechnical Commission, 1965).

1 MΩ is achievable in high quality metal film resistors.

Note that the IEC standard refers to the current-noise in a decade of bandwidth, but if you apply the noise to several decades of frequency you need to multiply by the square root of the number of decades. Thus root-decade is more helpful and more technically correct.

EX 4.8.1: A 100 kΩ resistor is on the limit of the above current noise spec. It is biassed with 100 μA from a relatively noise-free high impedance source. What is the **total** RMS noise across the resistor in the absence of any significant loading when looking at the noise in the frequency range from 0.001 Hz to 1 MHz.

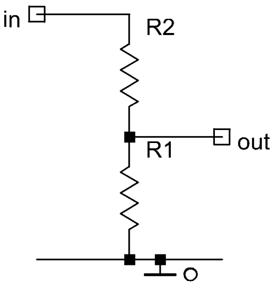
The standard measurement method is not without its problems however. The requirement for noise-free resistors has already been mentioned. Additionally, the test specifies the noise in a bandwidth of 1 kHz geometrically centred at 1 kHz. The system is then calibrated by injection of a 1 kHz signal. The impedance of the resistor under test at 1 kHz therefore affects the system calibration. This potentially causes problems for megohm value resistors having several picofarads of shunt capacitance. Such resistors would read lower on noise than they should.

In general for any type of component, flicker noise is indicative of a long term drift problem.

4.9 Preferred Resistor Ratios

When it comes to the routine action of designing a circuit using preferred resistance values, young designers are left on their own as far as their training is concerned. How do you decide what values to use?

FIGURE 4.9A:



Let's keep it simple and not worry about the source impedance driving 'in' and the load impedance on 'out'. All I want is a simple attenuation ratio. Let's say this is an attenuator operated at 10 kHz or less.

I can make a 2:1 attenuator by making the resistors equal. If I wanted a 3:1 attenuator, I could use three equal resistors with two in series for R2. For a 4:1 division ratio I could use four equal resistors in series, tapped-off on the bottom resistor. For a 5:1 ratio I could also use four equal resistors, with two in series at the top and two in parallel at the bottom. But in general, the ratio required could be any non-integer value; a method is needed to systematically find the values in a rapid manner.

You can pick one of the resistors to get the current through the chain around the right sort of value. The other value is then defined by the desired attenuation. Ordinarily one resistor will not fit the desired ratio exactly, so another resistor is then put in series or parallel to make the ratio closer.

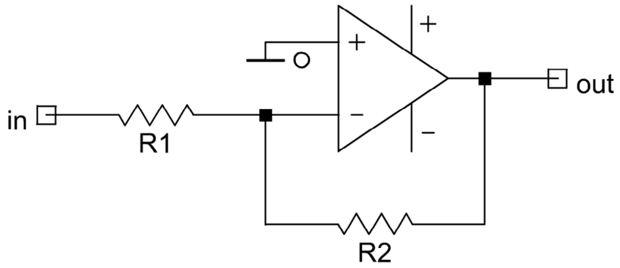
Suppose you want a division ratio of 7. Let's say that R1=1K because that gives a reasonable current and output impedance. In this case R2 is 6K0. If you are using E24 values then the closest you can get is 6K2. So you could use 6K2 and put 180K in parallel with it. This gives 5K994 which is pretty close. That was lucky. You might have had to try the next lower value and go for a series combination such as 5K6 and 390Ω (5K99). Or perhaps 5K1 and 910Ω (6K01). The trouble is that you have to try dozens of values to see if you actually have the best pair of resistors for the job. If the ratio is slightly off then you are wasting your tolerance margins.

It is convenient for a whole range of problems to reduce them down to the simple ratio of two resistors. For the attenuator problem, the desired division ratio is:

$$\boxed{T = \frac{R_1}{R_1 + R_2} = \frac{1}{1 + \frac{R_2}{R_1}}}$$

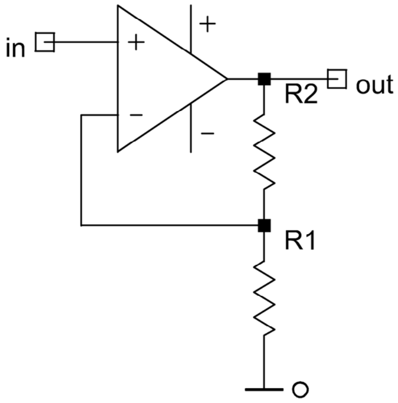
To get an attenuation of 7 you therefore look up a resistor ratio of 6 in a table.

FIGURE 4.9B:



For an inverting amplifier the gain is $\frac{R_2}{R_1}$.

FIGURE 4.9C:



For a non-inverting amplifier the gain is $1 + \frac{R_2}{R_1}$.

This is not something that you have to do every day, but it is useful to have a systematic way of solving the problem. The old answer was a chart of ratios using three resistors. The modern answer is a computer program.

The computer sorts through all the available ratios. If you want a ratio of 3.102 and a source impedance around 10K then the program will give you the closest ratios available. My version of this program is called *Selector*: **www.logbook.freeserve.co.uk**.

***EX 4.9.1:** You have a voltage reference of 1.250 V ±0.1%. You want to get a reference voltage of 1.090 V with the best accuracy possible, using at most three 1% resistors in the divider and no trimming. The load on the reference must not exceed 1 mA and the output impedance of the 1.09 V output must not exceed 10K. Consider this output as unloaded.

a) Select an optimum set of E24 values to give the required performance.
b) What are the resulting worst case initial limits on the attenuated reference output, expressed as voltages? (Neglect TC, noise, drift and the load on the reference.)

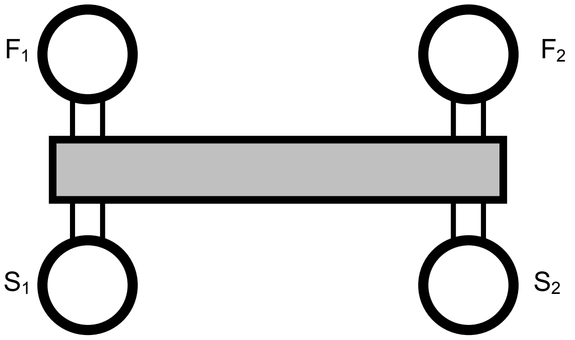
4.10 4-Terminal Resistors

If you have a resistor whose value is supposed to be 0.01 Ω, it should be evident that measuring the exact value of the resistance will be challenging. Any contamination on the terminals, or variation in contact pressure, could change the measured resistance by hundreds of percent. The same problems occur with 10 kΩ resistors, but then the error magnitude is reduced by 5 decades.

Precision measurement of resistance is done by injecting a known current and measuring the resulting voltage difference. The current (*force*) leads and the potential (*sense*) leads are kept separate all the way to the precision ohmmeter for best accuracy.

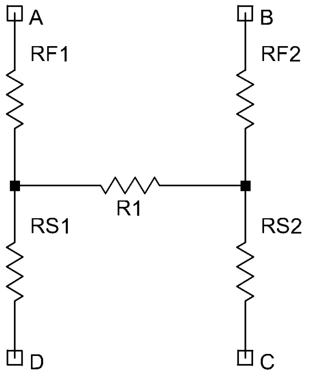
Figure 4.10A shows separate force and sense connections to a resistive element. The connections are made at hand-tightened screw thread terminals.

FIGURE 4.10A:



This plan view shows a rectangular resistive element (grey) connected to four terminals. The copper conductors are much deeper (into the page) than the resistive element to minimise their resistance. Copper is unsuitable for the resistive element itself because of its +0.4%/°C TC.

FIGURE 4.10B:



Given the physical construction of the resistor, the circuit model is 'obvious'.

In order to better explain the model, it is convenient to re-label the terminals F1=A, F2=B, S1=D, S2=C.

The individual elements in the model can all be directly measured using a precision 4-wire ohmmeter.

FORCE	SENSE	RESISTOR
AB	CD	R1
AD	AB	RF1
AD	CD	RS1
BC	AB	RF2
BC	CD	RS2

Typical measurement connections. You should realise that there is no unique connection for the measurement of any of the resistive elements in the model. Use the notation F(AB)S(CD) to mean force on terminals A & B, sense on terminals C & D.

R1= F(AB)S(CD) = F(AC)S(BD) = F(CD)S(AB) = F(BD)S(AC). And for each of these connection schemes either the force pair or the sense pair, or both, can be swapped without affecting the result. These equalities can be used to minimise thermal EMF errors. Notice also that F(AB)S(CD) = F(CD)S(AB); the force and sense terminals can apparently be interchanged without changing the measured value!

A minimum of 6 resistive elements are needed to model any general 4-terminal resistive device or network. The model given above has only 5 elements and is therefore incomplete! Nevertheless this 5 resistor model is widely known and used. Only a few specialist metrologists will even have heard of Searle's 8 resistor model.[15]

Searle's model is far from ideal for three reasons:

1) It uses 2 pairs of equal resistors to give 6 independent resistor values.
2) It is obtained indirectly by solving simultaneous equations.
3) Some of the resistive elements can turn out to be negative.

The 'fault' with the 5 resistor model is due to the finite value of F(AD)S(BC). If you look at the plan of the 4-terminal resistor you will see that current from F1 to S1 can 'spread' into the resistive element and there would be some slight potential difference resulting between terminals F2 and S2.

Assuming that the force / sense terminals and connections are equally rated for current, a novice would not interchange force and sense connections because of possible errors. An experienced, well-educated engineer might not bother about the force / sense connections, knowing they are equivalent. The true expert would take the same care as the novice! Interchanging force and sense terminals *can* introduce significant errors, but only on badly designed resistors measured very accurately.

Force terminals on low value resistors should be larger than the sense terminals. This indicates their function if there are no other markings. The force terminals are required to handle currents up to 12 orders of magnitude larger than those in the sense terminals!

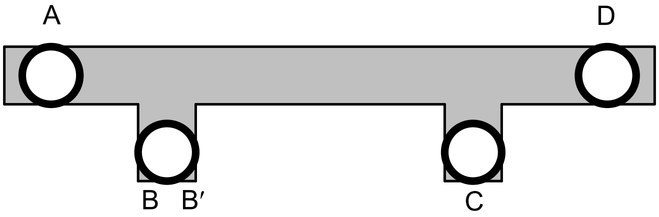
Having now designated the terminals as F1, S1, F2, S2, the resistance marked on the four terminal resistor is F(F1F2)S(S1S2). A useful design criterion for this resistor is the magnitude of the ratio $\left|\frac{F(\mathrm{F1F2})S(\mathrm{S1S2})}{F(\mathrm{F1S1})S(\mathrm{F2S2})}\right|$. This ratio is ideally infinite, but in 1911, when Searle presented his paper, resistors in the national metrology institutes had values for this ratio between unity and 10^{100}. Values of this ratio greater than 10^9 ensure that there is no significant error introduced by interchanging force and sense terminals.

The effect of a finite value of F(F1S1)S(F2S2) can be modelled by a resistor between either F1 and F2, or between F1 and S2. The position is chosen according to the polarity of the voltage measured in the F(F1S1)S(F2S2) test. It should be clear that if current is injected into F1 and removed from S1, F1 will be more positive than S1. If during this test F2 is more positive than S2 then the sixth resistor should be from F2 to F1. For a resistor designed like that in Figure 4.10A, the physical structure dictates that the sixth resistor is from F1 to F2.

[15] G.F.C. Searle, 'On Resistances with Current and Voltage Terminals', in *The Electrician*, LXVI (1911), pp. 999-1002, 1029-1033.

This 6 resistor model does not take into account *current spreading* in the resistor terminals themselves. Current spreading can be illustrated with an example.

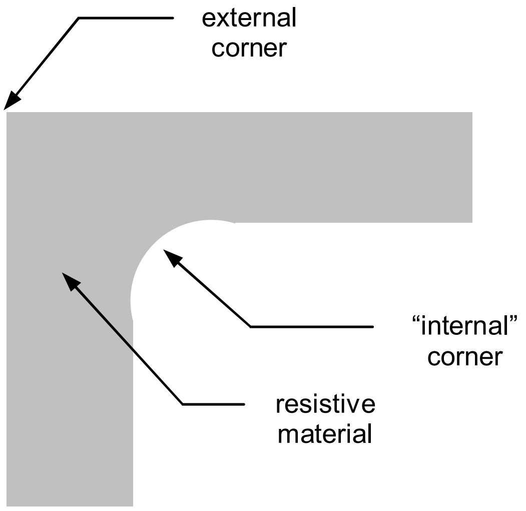
FIGURE 4.10C:



In this figure, the resistive element (grey) is a shaped sheet of a low TC metal such as Zeranin®. The circles are the terminals. If A and D are the force terminals, there will evidently be a significant voltage gradient across the face of the terminals B and C. The exact nature of the connection made at the sense terminals therefore has a significant effect on the measured resistance value. You can quantify this error source by measuring the voltage difference across the face of any terminal, say between B and B′, and comparing it to the voltage difference across the main resistive element.
If the voltage difference across the B terminal is less than 1 ppm of the voltage difference between B and C, the connection at B will not introduce a significant error to a 10 ppm measurement accuracy.

An improved design would extend the 'arms' out to terminals B and C, thereby minimising the voltage gradient across the terminal faces. Make the terminal extension arms greater than 3× the conductor width and the problem will be removed.

FIGURE 4.10D:



Always round the "internal corners" in a resistive element. This keeps the current density more uniform, reducing hotspots in the resistive material. Internal corners disrupt the current flow the most.

External corners need not be rounded for operation at low frequencies or where stray capacitance is not a problem. For RF, microwave, mm-wave or high impedance attenuator applications, it is preferable to also round or bevel the external corners in order to minimise stray capacitance.

® Registered trade mark of Isabellenhütte.

CH5: the potentiometer

5.1 Types

The potentiometer was devised by Poggendorff in 1841 as an "accurate" means of comparing voltages.[1] This application still survives today in forms such as the ***Kelvin-Varley*** divider, and the binary voltage divider.[2] Typical usage of the potentiometer, however, is as a variable resistor. This is unfortunate because the component gives more accurate performance when used correctly.

Potentiometer, variable resistor, pre-set potentiometer, preset, pot, Trimpot®, trimmer and rheostat all vaguely refer to the same type of component. A wiper slides down or around a conductive track, giving a three terminal component. The term *rheostat* refers to a wire-wound high power device.

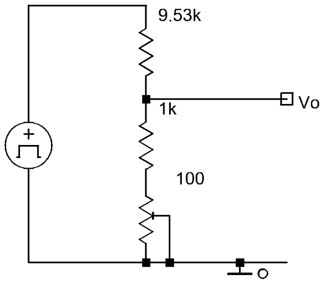
This chapter is devoted to the *preset potentiometer*, used for internal adjustments within equipment. I will call this a *pot* for simplicity, although the terms *preset* and *trimmer* are also widely used in industry.

Note that the term *trimmer* is also used for preset variable capacitors. The term "preset" is used to indicate an internal trimming or calibration component, as compared with a user-operated control.

For audio systems the standard volume control has a non-linear characteristic optimised for audio systems. This characteristic is specified as a *logarithmic control law* or an *audio taper*. Control potentiometers are therefore specified as log, lin, or anti-log. For a fixed DC input voltage across the track, a lin pot gives a wiper voltage which changes linearly with rotational angle. A log pot output starts slowly then speeds up dramatically. Half way around the output might only be 10% of the full output. An anti-log characteristic is rotationally reversed compared to a log characteristic. In this chapter only lin characteristics are discussed.

5.2 Temperature and Time Stability

FIGURE 5.2A:



Here is an example of an all too common error. This circuit is intended to be a precision 10:1 attenuator. The pot has been wired-up as a variable resistor. Whilst the fixed resistors are good 1% 10 ppm types, the pot is a 20% 500 ppm type. The (long term) load-life stability of the pot is given as 10%.

***EX 5.2.1**: Neglecting the source impedance, the load impedance and any effective rotation of the pot, calculate the gain TC and the long term drift of the gain for both extreme positions of the pot. Use 0.5% drift for the fixed resistors and 5% for the pot.

[1] J.A. Fleming, in *Fifty Years of Electricity: The Memories of an Electrical Engineer* (The Wireless Press Ltd, June 1921), pp. 274.

[2] S. Hoi Tsao, 'A 25-Bit Reference Resistive Voltage Divider', in *IEEE Transactions on Instrumentation and Measurement*, IM-36, no. 2 (June 1987), pp. 285-290.

® Trimpot is the registered trade mark of Bourns, Inc.

The pot is unnecessarily ruining the performance of the attenuator. You could pay more and get a better pot, but you could also get a better result by changing the wiring of the circuit.

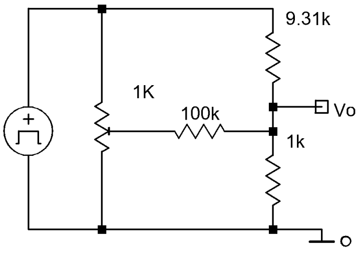
Data sheets refer to the *end resistance* of the pot as being between 1% and 3% of the track resistance at each end of the travel. This is the minimum resistance between the wiper and the track connection at each end of the travel.

EX 5.2.2: What is the *guaranteed* minimum gain adjustment range of the previous circuit, using 2% end-zones at each end of the pot and a 3 Ω wiper resistance?

Some mistaken individuals omit R2 completely, thereby "saving" a resistor. This is a ridiculous thing to do, as it gives 10× the error and 10× less adjustability. Never use a pot for 100% of a resistance when you intend to trim it by say <10% unless accuracy is of no importance in that design.

A pot should be wired as a *potentiometer* and not as a variable resistor whenever possible. The added cost for this is usually the inclusion of one extra resistor, but the improvements to circuit stability are well worth it.

FIGURE 5.2B:



Apply the same rules as before:

The main attenuator resistors are 1% 10 ppm and with 0.5% drift. The pot is 20% 500 ppm and has 5% drift.

The 100K resistor is a cheap 5% 100 ppm component with 1% drift.

***EX 5.2.3**: Neglecting the source impedance, the load impedance, the end-resistance of the pot and any effective rotation of the pot, calculate the gain TC and the long term drift of the gain for both extreme positions of the pot.

EX 5.2.4: Neglecting the pot end-resistance,

a) What is the guaranteed gain adjustment range of the above attenuator?

b) Why is the gain adjustment range particularly important for these exercises?

From these exercises you should have seen that the potentiometer configuration had half the TC of the variable resistor configuration. It also shows that the 100K resistor TC is having too much effect; a ±50 ppm/°C part is definitely preferable here.

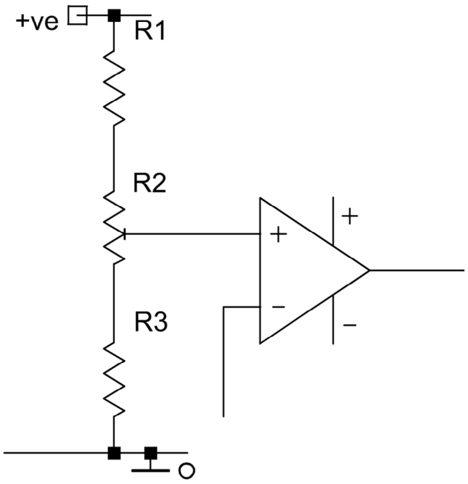
These exercises demonstrate a very real problem and a very useful solution. You can buy better pots, but the point of the exercises was to show you how to get the best performance from whatever pots you actually use. Also, if you use a pot which is 10× as expensive in order to 'save' one cheap resistor, you will not be very popular with your employer.

There is only one problem with the way the potentiometer has been wired up in the previous attenuator example; it puts additional load on the input to the attenuator. This arrangement is excellent for attenuators used inside equipment, but it is often unsuitable for an attenuator that interfaces to the outside world.

Let's suppose that you are making a high-voltage input attenuator, designed for several hundred volts or more. This would be a very nasty place onto which to connect the track of a pot. In this case you would have to revert to some sort of variable resistance configuration and you might have to pay more money for a better part.

If you use carbon pots then a TC of 1000 ppm is not unusual. ***Cermet*** types are what would normally be used. Their TCs are in the range of 50 ppm to 200 ppm. Below 20 ppm you need to use expensive material like Nichrome and you start paying significantly more for it, say 5× the price.

FIGURE 5.2C:



I hope I have convinced you that a pot should be wired up as a potentiometer and not as a variable resistor. In this circuit it may appear as if the pot is wired up correctly as a potentiometer; it is not. The track resistance is uncertain and relatively unstable. Because of the way it has been connected, the voltage across the track is not stable with time and temperature. Thus the output voltage is not stable with time and temperature.

***EX 5.2.5:** In the above circuit R1=3K3, R2=1K, R3=3K3. For the purpose of this exercise, neglect the selection tolerances on all the components. Also neglect the TC of R1 and R3 [since they are much less than the TC of the pot]. The TC of R2 is 500 ppm/°C. Assume that the +ve supply is totally stable at exactly +10 V. What is the worst case TC on the output [wiper] of the pot?

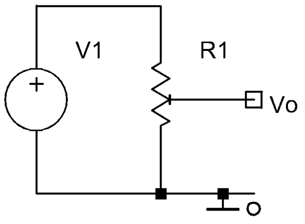
5.3 Analysis

In order to analyse pot configurations, I am going to use the *Per-unit Potentiometer Position*, symbol *p*. (In Maxwell's *Treatise on Electricity & Magnetism* [1891], *m* is used. Some recent texts use *k*.) I will also express *p* in percentage terms.

p is the effective rotation of the pot: 0 represents no rotation, 1 represents full rotation. The wiper at the grounded end of the pot is normally the 0 position.

End-resistance does not seem to affect the range of *p*; $0\% < p < 100\%$ is easy to achieve. Nevertheless, the end-zone characteristic may not be as regular as the main track characteristic. Consider the useable range as $1\% \le p \le 99\%$ unless you have information to the contrary.

FIGURE 5.3A:



In this situation $V_O = pV_1$. The output resistance is 0 at the top and bottom positions of the pot. The maximum output resistance is *R/4* and occurs at the mid-point. The 'bottom part' of the resistance is pR and the 'top part' is $(1-p)R$, the output resistance being these two resistances in parallel.

Using the 'product over sum' rule for parallel resistances:

$$R_O = \frac{pR \times (1-p)R}{pR + (1-p)R} = R^2 \cdot \frac{p(1-p)}{R} = pR(1-p)$$

EX 5.3.1: Prove that R_O (max) is *R/4*.

Answers to the following questions are needed for each pot used:

- With what resolution can *p* be set initially?
- What is the spec for the variation of *p* with time?
- What is the spec for the variation of *p* with temperature?
- What other factors affect *p*?

Unfortunately manufacturers of pots do not give these figures, apart from the resolution figure; this is often given as infinite! The aforementioned specs are completely missing from data sheets. This being the case, there is no choice but to invent our own numbers and use them.

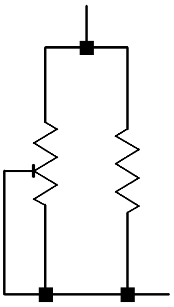
Analog engineers consider it a point of pride to be able to adjust any cheap & nasty pot to an exact position. You can effectively rotate the pot slightly by tapping the pot with the trimming tool or applying a light impulsive pressure. This is a physical skill learned by experience and every analog engineer has his/her own little tricks.

Some pots feel gritty and others feel smooth. The smooth ones are easier to adjust to a high resolution. How finely should a designer require an operator to set a pot? The manufacturer says you can set it to infinite resolution: rubbish! Take a pot and wire it across a 10 V supply from a DC calibrator, then measure the wiper voltage with a 5½ digit DVM. Do you think you can set the voltage to within a digit? If so then you should definitely try it!

Beyond a certain resolution it takes too long to set the pot to the required value. We therefore chose a value so even inexperienced technicians can set the pot quickly. The number chosen is 0.2% (p=0.002). On the 10 V test, the pot has to be set to within 20 mV. A good pot would be at least twice as good as that.

***EX 5.3.2:** A poorly trained senior engineer has designed a front-end amplifier using an opamp with a maximum offset voltage of ±7 mV. You point out that this needs to be corrected as you are measuring very small signals. He comes back having put in a ±10 mV correction using a single-turn pot. He says this can easily be adjusted down to the ±10 μV that the system requires. Comment *in depth*.

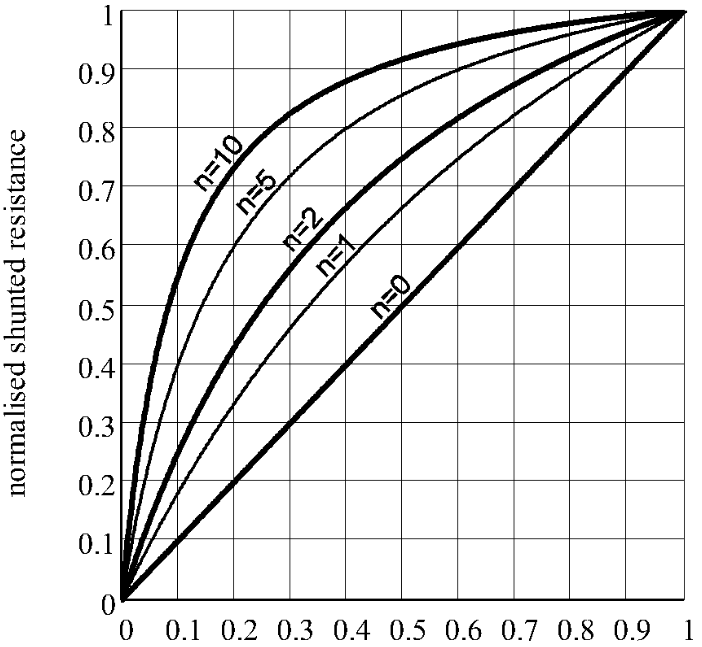
FIGURE 5.3B:



The Potentiometer Handbook from Bourns Inc [dated 1975] talks about the "adjustability" of a pot. They evaluated this as the error on a midway setting of the pot achieved within 20 seconds.

Using *p* as an analysis tool, it is now possible to consider the effect of shunting a variable resistor with a fixed resistor. The idea is to reduce the effects of drift and TC in the pot's track resistance.

FIGURE 5.3C:



These curves show the effect of paralleling the pot by a fixed resistor. Curve 0 is with no shunt resistor. Curve 1 has a shunt equal to the pot track. Curve 2 has a shunt ½× the track resistance. This continues up to curve 10 where the shunt is 1/10th the track resistance. Notice how the curves become more non-linear as the shunt is decreased. The full control range is still present, but the rotational sensitivity is worse. In trying to make the circuit more stable, you could make it less stable.

This scheme is poor because the pot is being used as a variable resistor.

5.4 Stability of p

Clearly vibration is a major concern to the long term stability of *p*. In situations where equipment is subjected to repeated or continuous vibration, it is unsafe to use pots without extra protection. The vibration could be sufficient to rotate the pot an arbitrary amount. Typical solutions to this problem include:

- Using "factory selected" {set on test} fixed resistor values instead of pots.
- Using PCB mounting DIP / rotary switches to switch-in fixed resistors.
- Cutting wire links which are shorting-out series connected resistors.
- Using solder-links across special pads to connect shunt resistors into circuit.
- Using microprocessor controlled DACs / switches / digital pots.
- Sealing the pot with a small dab of paint on the cover. [Finger-nail polish and car touch-up paint are convenient as both come with a built-in brush.]
- Melting the rotating part of the pot to the body using a swift touch of a soldering iron. [This *historic* technique is no longer recommended.]

Some of these solutions should only be necessary if the equipment is subjected to continuous vibration. Otherwise, vibration in transit should not be so great as to cause a problem if the pot has a suitable adjustment range.

The digital pot and DAC solutions have become popular because prices have dropped and they allow adjustment by less skilled staff. They also allow computerised adjustment. Additionally they allow "covers-on" calibration. Putting covers onto an instrument *always* changes the calibration, although sometimes the effect is not too great.

Historically equipment had access holes to allow adjustment of trimmers without the need to remove the covers. This is not done very often in modern designs. The modern approach is to prevent users adjusting calibration constants. Computer controlled calibration schemes should therefore be "locked out" to everyone except the calibration and service personnel.

To test the stability of p, I did a little experiment with some inexpensive Citec 406 20K cermet pots [circa 1999]. The test was to wire up 8 pots across a common supply and measure the wiper voltage, after having set them roughly by eye to mid-range. Then they were left to rattle around on the dashboard of my car for 3 days and nights during a mild UK Spring. They travelled about 30 miles during this period. The voltage supply was a precision DC calibrator with 1 ppm resolution and they were measured using a 5½ digit DVM with >10 GΩ input resistance.

The calibrator was adjusted to make the DVM read exactly 10.0000V before both sets of readings; thus only short-term stability of the DVM was required. These are un-adjusted, raw figures with none omitted. Pots 1-4 were wired clockwise and pots 5-8 were wired anti-clockwise.

POT NUMBER	Set Point @ 26.5°C	Later Value @ 25.2°C	Delta p (%)
1	5.1549 V	5.1517 V	–0.032
2	4.6047 V	4.6053 V	+0.006
3	5.5256 V	5.5250 V	–0.006
4	5.1182 V	5.1178 V	–0.004
5	5.0834 V	5.0807 V	+0.027
6	4.8725 V	4.8732 V	+0.007
7	4.9613 V	4.9641 V	+0.028
8	4.6453 V	4.6455 V	+0.002

This is just a little experimental data to justify putting a genuine figure on the value of Δp. Frankly I was rather surprised the results were so good, despite the pots not being locked by any of the anti-vibration methods mentioned.

How about temperature? You should expect that the track of a pot is fairly uniform. The thickness of the resistive element should be fairly consistent and you should expect that the sections of track either side of the wiper would have TCs that were matched to better than 10:1 of their ±100ppm/°C spec. So I did a test on this.

All pots had been moved since the previous test. I measured the pots, heated them with a hairdryer for a few minutes and measured them again. To detect any permanent shifts, I let them cool down and measured them again.

Pot Number	22°C Ref	45°C	from ref Δp (%)	20°C	from ref Δp (%)
1	4.7708	4.7717	+0.009	4.7704	–0.004
2	4.7739	4.7742	+0.003	4.7731	–0.008
3	5.3488	5.3517	+0.029	5.3518	+0.030
4	5.2923	5.2925	+0.002	5.2913	–0.010
5	4.5821	4.5813	–0.008	4.5832	+0.011
6	4.6077	4.6077	0.000	4.6079	+0.002
7	5.5271	5.5257	–0.014	5.5272	+0.001
8	4.9944	4.9952	+0.008	4.9951	+0.007

The designer needs to have a figure for Δp to work with. This should ideally come from the manufacturer; failing that, experiment or experience has to suffice. In the absence of those, use the value, Δp= 0.005 (0.5%) for single turn pots and Δp= 0.002 (0.2%) for multi-turn pots. These apply to cermet pots (or better), but not to carbon pots. I have not been 'generous' with the multi-turn pot spec; although they should be more mechanically stable, the conductive track can be the same length as a single turn track. For this reason the TC and drift tracking of the resistive element may not be much better. Don't blame me if your pots are worse than this; I can't be held responsible for your use of sub-standard components!

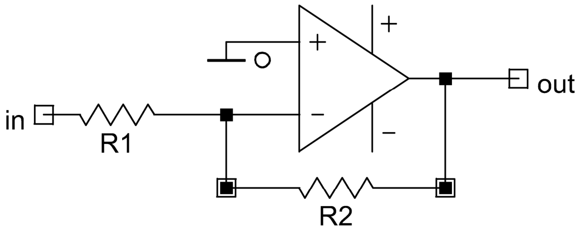
To get the best out of pots, follow these four simple rules:

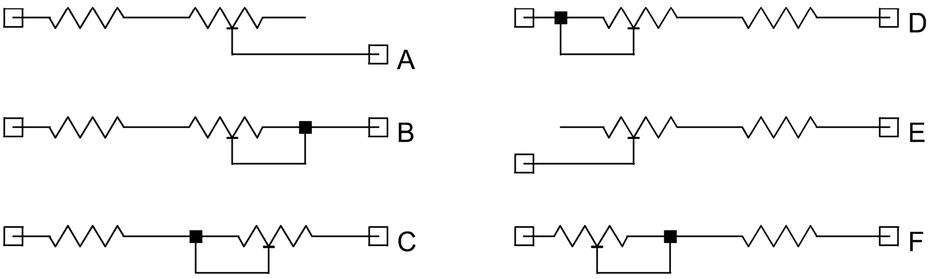
- Wire a pot as a pot, not a variable resistor, whenever you can.
- Give the pot as little adjustment range (of your circuit) as possible
- Always join all three terminals of a pot to your circuit.
- When used as a variable resistor, connect the wiper and one end of the pot to the lower impedance point or the less sensitive point in your circuit.

Opamps are covered in a later chapter (CH10), so it may be best to skip over this next exercise and come back to it later.

***EX 5.4.1**: Here are some circuit configurations for an opamp gain control. They are all wrong since they are using the pot as a variable resistor.

FIGURE 5.4A:



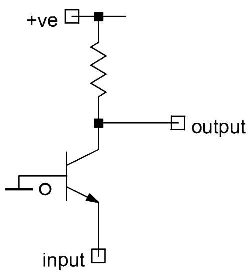


Each of the networks shown could be used (one at a time) to replace R2 in the orientation shown. Which configuration(s), if any, are better and why?

The most likely failure mechanism on a pot is for the wiper to go open-circuit. Connecting the third pin when using the pot as a variable resistor not only reduces the rotational noise when trimming, it also limits the possible error if the wiper should go open-circuit in service.

EX 5.4.2: This circuit is supposed to represent a simple *common-base* video driver. For the sake of the example, suppose that 10% adjustment range is needed on the resistance.

How should the pot be wired up? **FIGURE 5.4B**:



The last two examples should have demonstrated the reason for wiring the pot to a low impedance point in the circuit. There is an exception to this rule, however. On a high voltage output amplifier, wiring the pot to the high voltage makes adjustment slightly more dangerous for an operator, particularly if the pot has an 'open frame' construction.

What I have not fully justified is the rule about tying all three pins of the pot into your circuit. I have mentioned the open-circuit wiper as a failure mechanism, but there is another reason. The wiper has a finite resistance which depends on the way it is made. Manufacturer's quote a 'wiper resistance' of at least 3 Ω, but this may not seem like much of a problem to you at the moment.

EX 5.4.3: A *junior* engineer has wired up the circuit shown here. The pot is a wire-ended type and the capacitance from the 'unused' pin to a nearby digital line is 0.2 pF. This digital line is 4000B-series CMOS, swinging 12 V ptp at 1 kHz. Draw an equivalent circuit for the pickup mechanism, setting the pot at mid-travel, so the interference can be simulated.

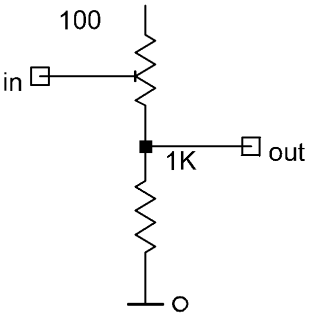


FIGURE 5.4C:

Assume that the input is driven from a low impedance source (<0.1 Ω).

There is a more general rule of interconnections which you may not have spotted in previous examples; let me spell it out for you. In elementary texts the order in which components are placed in a series chain is not important. In a real-world circuit this order of connection can make all the difference between a working circuit and a non-working circuit. *Always* consider the stray resistances and capacitances on the connecting nodes in any series chain.

Experienced engineers know that the act of re-laying a PCB, doing something simple like putting one component in parallel with another, can wreck an otherwise working design when the circuit is working above say 10 MHz. Surface mount 1206 parts can often be placed on top of each other, and this may be a better long term solution than trying to move everything else out of the way in order to "correctly" place the extra component on the PCB. In any case, on such a change it is wise to be cautious and not build hundreds of the next issue of the PCB before a few have been tried.

When laying out a PCB it is good practice to define the pin connections such that the direction of rotation of the pot is taken into account. If the pot adjusts a regulator output voltage, for example, it is useful to define clockwise adjustment as always increasing the output voltage.

One school of thought suggests that it is important to pass a significant current [>25 μA] through the wiper of a pot. This reduces problems with ***dry circuit*** conditions in the wiper contact resistance. I do not bother about this particular issue. The point is that the contact resistance may well go "out of spec" in the sense of being greater than 3 Ω or some such value. If it goes up to 100 Ω, but is in series with 100 kΩ, this is not going to cause a problem. Also, it is quite likely that the voltage will exceed the 100 mV level, which should puncture any contaminant build-up and restore the wiper resistance to its correct range.

CH6: the capacitor

6.1 Basics

A *capacitor* is an electrical component having *capacitance* as its dominant electrical attribute The obsolete term *condenser* is rarely heard, except for older automotive applications. All of the same factors that adversely affect resistors also affect capacitors to some degree, but the emphasis depends very much on the type of capacitor. Whereas resistors use resistive materials, the key to a capacitor is its *dielectric*.

The application of an electric field to an insulating material, a dielectric, stores both charge and energy. A dielectric always increases the *electric flux density*, **D**, above that you would get in free space; the factor of increase is called the *dielectric constant* by some authors and the *relative permittivity* by others. A physicist, for example, would write $\mathbf{D} = \varepsilon\,\mathbf{E}$, relating the electric flux density **D** to the electric field intensity **E** by the permittivity ε. This relationship can also be expressed as $\mathbf{D} = \varepsilon_0 \varepsilon_r \mathbf{E}$, where ε_0 is the permittivity of free space, 8.85 pF/m, and ε_r is the (dimensionless) relative permittivity which is unity for free space.

The plates in a capacitor attract each due to their storage of opposite charges. Thus a higher applied voltage necessarily gives a higher attractive force and therefore a very slightly smaller gap between the plates. At least some voltage coefficient of capacitance should therefore be expected, depending on the rigidity of the dielectric.

The lowest value of dielectric constant you can get is 1.0000 for free space, with air coming in a very close second at 1.0005.

The terms *dielectric constant* and *relative permittivity* are synonymous and interchangeable. The synonymous term *relative dielectric constant* has also appeared in text books from time to time, but is an incorrect variant. In texts prior to 1940, the term *specific inductive capacity* is used to mean relative permittivity.

If you need to minimise stray capacitance, there are five possible methods.

i. Increase the spacing
ii. Reduce the amount of conductor present
iii. Reduce the amount of dielectric present (use more air)
iv. Decrease the dielectric constant
v. Use a ***bootstrapped*** guard/shield.

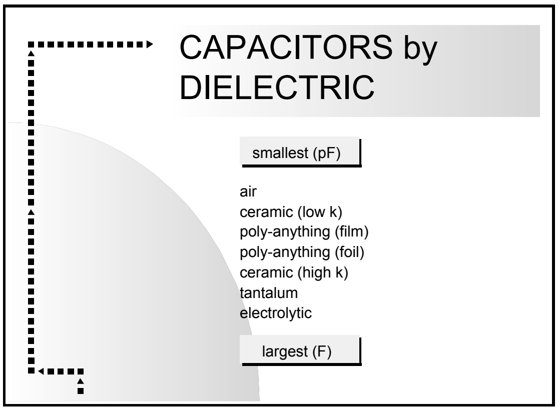
One way of removing dielectric on a PCB is to machine away as much of the unnecessary PCB material as possible in the critical area. Standard FR4 PCB material has a dielectric constant of around 4, so removing this material can reduce the overall capacitance significantly.

Reducing the amount of conductor present might be achieved by not using a plated-through hole on a double-sided PCB or by removing unnecessary pads on wire-ended component positions (both just in key circuit positions).

The use of a bootstrapped guard or shield requires an amplifier, and this topic will be explained fully in a later chapter. For now, just think of alternating current flow in the stray capacitance. The alternating current is proportional to both the capacitance and the (sinusoidal) voltage across the capacitance. Reducing the capacitive current can be done directly, by reducing the capacitance, or indirectly by making the nearby metallic objects at the same potential as the sensitive region.

The first thing to do is to list the available types of capacitor. Since it is the dielectric which governs the behaviour of a capacitor, capacitors are specified in terms of their dielectric more than anything else.

FIGURE 6.1A:



This figure is grossly over-simplified and the bands of capacitance are not as clear-cut as the figure would suggest. By poly-*anything* is meant polyester, polycarbonate, polypropylene, poly-phenylene sulphide, polysulfone, poly-styrene, … in fact poly-anything!

Air is an important dielectric for capacitors, not least of which is because it is very nearly ideal in terms of linearity, dielectric loss &c. It can also give rise to significant unintentional capacitance for large structures or long cables. If you have a screened box near to an earthed plate, for example, the capacitance can be an important consideration.

Even laboratory standard precision air capacitors are significantly affected by humidity and barometric pressure. A 30 ppm shift could occur for a 15% change in relative humidity or for a 50 millibar[†] change in atmospheric pressure for example.

The simple formula for a parallel plate capacitor neglects the fringing field [edge-effects] and therefore always gives a low answer. For this formula to be accurate the plate separation has to be more than 100× less than the smallest linear dimension of the plates.

$$\boxed{C = \frac{\varepsilon_o \, \varepsilon_r \, A}{h}} \qquad \varepsilon_0 \cong 8.85\,\text{pF/m}\,.$$

A is the plate area in square metres, *h* is the plate separation in metres. In capacitor terminology, ε_r is usually called the *dielectric constant*, κ.

EX 6.1.1: A screened amplifier has one side 5 mm from an earthed chassis plate. The amplifier screen is 80 mm ×170 mm. Neglect the fringing capacitance, treating the amplifier's case as a simple parallel plate capacitor. What is the capacitance between the amplifier screen and the chassis plate?

The parallel plate capacitance formula is widely overused in situations where it is grossly in error. Unfortunately even specialist electromagnetics text books do not give formulae for the fringing field correction factor.

[†] 1000 millibar = 1 bar, approximately one standard atmosphere.

If the capacitor is long compared to its width the fringing at the ends becomes less important and it is possible to consider the field as being approximately two-dimensional. Now *conformal mapping* can be used to determine the field lines exactly.[1] Unfortunately the apparently neat solution only gives an *implicit* answer in terms of ***elliptic integrals***. Numerical iteration is then required to extract an answer to any specific problem.

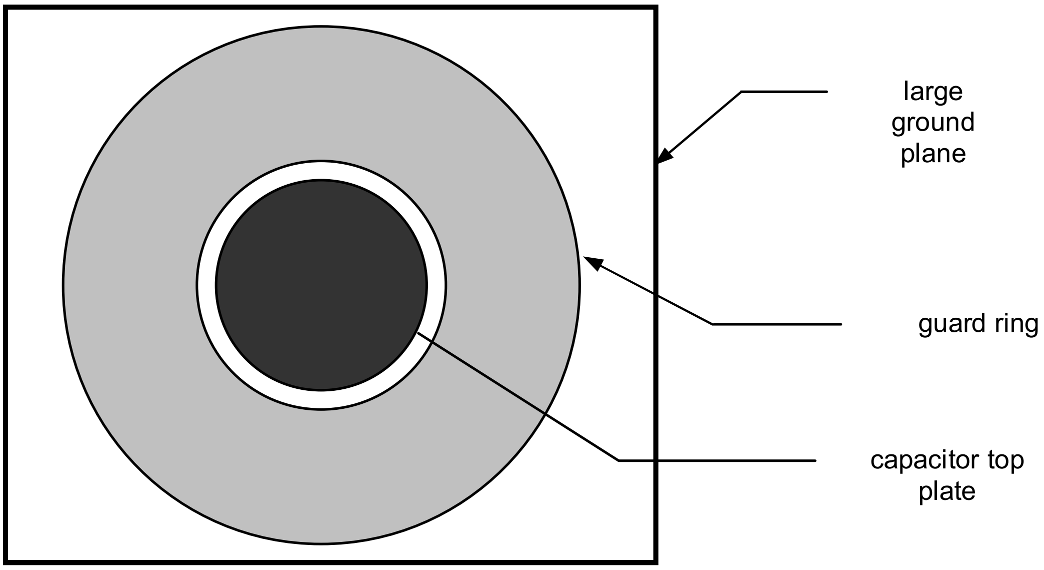
$$\boxed{C = \frac{\varepsilon_0\,\varepsilon_r\,A}{h}\cdot\left[1+\frac{h}{\pi w}\left(1+\ln\left(\frac{2\pi w}{h}\right)\right)\right]}$$

This useful approximation to the exact 2D field dates back to Kirchhoff (1877).
w is the plate width, $w > h$

Note that when *w*=*h* the required correction factor is 2.1×, a 110% increase due to the fringing field!

For precision calculable air-dielectric capacitors, the problem of the fringing field was solved by the use of Kelvin's *guard ring*. Imagine a large ground plane above which is a large parallel metal plate. The field between the plate and the ground plane gets arbitrarily uniform with sufficient distance from the edges. If the top plate is cut to make a smaller inner plate, electrically separate from the outer remaining plate, but held at the same potential, the electric field can still be reasonably uniform, provided that the cut is very thin. The fringing field still exists outside of the guard ring, but is not relevant to the value of the capacitor.

FIGURE 6.1B: PLAN VIEW



The "guard ring" and the "capacitor top plate" in the above drawing are both metal plates, and are both in the same plane, parallel to the ground plane.

There is capacitance *to* something, usually to ground {earth}. It is not necessary for that something to be nearby. Even an "isolated" wire has a capacitance to ground.

[1] A.E.H. Love, 'Some Electrostatic Distributions in Two Dimensions', in *Proceedings of the London Mathematical Society*, Series 2, vol 22 (1923), pp. 337-369.

$$\boxed{C = \frac{2\pi\varepsilon_0\varepsilon_R}{\ln\left(\frac{r_{\text{ext}}}{r_{\text{int}}}\right)}\ \text{F/m}}$$

This is the capacitance formula for a standard coaxial cable. Measurements are made from the conducting surfaces which are closest to each other. r_{int} is the outer radius of the inner conductor and r_{ext} is the inner radius of the outer conductor.

This is the capacitance formula for a round inner conductor covered by a coaxial square outer conductor, where d is the outer diameter of the inner conductor, and D is the inner side of the outer square conductor.[2]

$$\boxed{C = \frac{2\pi\varepsilon_0\varepsilon_R}{\ln\left(\frac{D}{d}\right) + 0.0704}\ \text{F/m}}$$

Wire inside equipment is unlikely to be more than 10 cm from other wires or the chassis metalwork. The minimum capacitance of such a wire can therefore be estimated by taking r_{ext} as 10 cm and r_{int} as 0.1 cm. This gives a capacitance of 12 pF/m. Increasing the external radius to 1 m only reduces the capacitance to 8 pF/m, whereas reducing the radius to 1 cm increases the capacitance to 24 pF/m.

$$\boxed{C = \frac{\pi\varepsilon_0\varepsilon_R}{\text{arccosh}\left(\frac{h}{d}\right)}\ \text{F/m}}$$

The capacitance formula for parallel wires, where h is the distance between the centres of the two wires and d is the diameter of the conductor. Using a 10 cm separation and a 1 mm diameter conductor gives 5.2 pF/m.

When $h > 5d$ the capacitance can be written as … with less than 0.5% error. This formula better illustrates the logarithmic behaviour of the capacitance with separation.

$$\boxed{C = \frac{\pi\varepsilon_0\varepsilon_R}{\ln\left(\frac{2h}{d}\right)}\ \text{F/m}}$$

Having been given the accurate formula for a pair of parallel cylinders above, it is a simple matter to consider the (equipotential) plane of symmetry between the two conductors as a conductor using the *method of images*.

EX 6.1.2: What is the formula for the capacitance of a conducting cylinder of diameter d separated from a ground plane by a distance s and parallel to the ground plane *Warning: It is not the axial separation of the cylinder from the ground plane that is being used here.*

Calculation of the capacitance of irregular, non-symmetrical configurations is remarkably difficult. For this reason it makes sense to try a few known configurations and see how much agreement there is between the results. This at least gives some sort of limit to the problem at hand.

Taking the formula for the capacitance of concentric spheres, with the outer sphere at infinity, gives a minimum value of the capacitance of any sphere (diameter d).

$\boxed{C \geq 2\pi\varepsilon_0 d}$ giving ≥ 0.556 pF for a 1 cm diameter sphere

[2] S.A. Schelkunoff, 'Conformal Transformations', in Applied Mathematics for Scientists and Engineers, 2nd edn (D. Van Nostrand Company (1948) 1965), pp. 298-300.

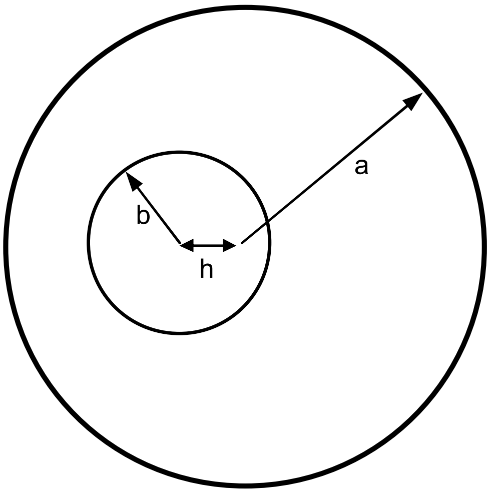
A circular disk has a slightly lower capacitance to infinity (diameter d):

$\boxed{C \geq 4\varepsilon_0 d}$ giving ≥ 0.354 pF for a 1 cm diameter circular disk.

Below is the formula for the capacitance of a vertical wire (height h, diameter d) to a horizontal ground plane. The formula is a fair approximation, even if as much as the top 2/3$^{\text{rds}}$ of the wire is bent over parallel to the ground plane. The formula also works if the whole wire is bent at an angle of up to 45° relative to the perpendicular. The capacitance is evidently most strongly associated with the total length of the wire, provided the bends do not fold the wire back on itself.

$$\boxed{C = \frac{2\pi\varepsilon_0 h}{\ln\left(\frac{1.155 \times h}{d}\right)} \approx \frac{0.56}{\ln\left(1.16 \times \frac{length}{diameter}\right)} \text{pF/cm}} \quad \text{for } \frac{length}{diameter} > 10$$

FIGURE 6.1C:



For the apparently very difficult problem of the capacitance between non-concentric cylinders, a relatively simple formula can be written:

$$\boxed{C = \frac{2\pi\varepsilon_0\varepsilon_r}{\operatorname{arccosh}\left(\frac{a^2 + b^2 - h^2}{2 \cdot a \cdot b}\right)} \text{F/m}}$$

This formula is useful for evaluating the change in capacitance of nominally coaxial conductors subject to manufacturing tolerances.

6.2 Dielectric Constant

Pull out your electromagnetics/physics texts to remind yourself of the fundamentals. The dielectric increases the *electric flux density*, symbolically expressed as $\mathbf{D}$. This increase is understandable when you think of the relation $\mathbf{D} = \varepsilon \times \mathbf{E}$. For a given potential difference between two conductors, increasing the relative permittivity gives more electric flux density and therefore more stored charge. The capacitance has increased, as more charge is stored for a given potential difference: $Q = C \times V$.

When there is a discontinuity in the dielectric in the direction of the electric flux density, the electric flux will be continuous across the boundary, and therefore the electric flux density will also be continuous. However, the electric field intensity **E** will be discontinuous at a distinct boundary between two dielectrics having different relative permittivities.

@EX 6.2.1: Two large parallel conducting plates are separated by air, but with a large flat plastic sheet lying against one of the plates. The dielectric constant of the plastic is K. The sheet fills a proportion, G, of the total gap. Where G=0 would mean no plastic and G=1 would mean all plastic. Derive the modification to the parallel plate capacitance formula, showing the physical reasoning behind the derivation. Neglect the fringing field.

Now you should be able to extend the idea, presented in the previous example, to the case of cylindrical symmetry, by considering a wire with an insulating sheath.

EX 6.2.2: What is the capacitance per metre of a 1 mm diameter copper wire covered by a 1 mm thick plastic sleeve of dielectric constant 5? For the purposes of calculation, assume the wire is surrounded by a (coaxial) 100 mm diameter earthed pipe.

Adding dielectric cladding to any conductor will always increase the capacitance and the RF power loss. Thus for low-loss coaxial RF-feeders, it is usual to remove most of the dielectric by using a foamed plastic (lots of trapped air bubbles), supporting the inner conductor on posts every few centimetres, or running the support as a spiral of insulating material. Note that even for good dielectrics, such as Polystyrene and PTFE, the conductivity of the dielectric increases proportionately to frequency to a good approximation. Thus whilst the dielectric loss can be minimal in the kilohertz region, it is a million times worse in the gigahertz region.

For an unspecified material between two parallel conducting plates, imagine an equivalent circuit in terms of a resistor and capacitor in parallel. As the frequency increases, the current becomes more and more reactive. A useful index of a material is therefore given by the ratio of the resistive to reactive currents.

$$dissipation\ factor = \frac{V}{R} \times \frac{1/\omega C}{V} = \frac{1}{\omega CR}$$

This is a lumped-element model rather than the material properties. The resistance for a parallel plate element is given by $R = \frac{\rho d}{A}$, where d is the plate separation and A is the plate area. Likewise for the capacitance, $C = \frac{\varepsilon A}{d}$. Thus

$$dissipation\ factor = \frac{1}{\omega CR} = \frac{1}{\omega \times \frac{\varepsilon A}{d} \times \frac{\rho d}{A}} = \frac{1}{\omega \varepsilon \rho}$$

This equation is usually written using the conductivity rather than the resistivity, giving

$$dissipation\ factor = \frac{\sigma}{\omega \varepsilon}$$

$\frac{\sigma}{\omega \varepsilon} > 100$ … a conductor, as the reactive current is less than 1% of the total.

$\frac{\sigma}{\omega \varepsilon} < \frac{1}{100}$ … a dielectric, as the resistive current is less than 1% of the total.

These definitions also apply to a propagating electromagnetic wave. Any material therefore behaves as a dielectric at a high enough frequency.

Looking more deeply at the subject of dielectric constant, the origin of the apparently increased electric flux is seen to be *polarisation* within the dielectric.[3] This polarisation takes many different forms according to the nature of the dielectric. At a molecular level there are electron shells surrounding charged nuclei, and there is some sort of bonding of one atom to another based on shared electrons. At this level there is an asymmetry of electric charge and therefore an external electric field can stretch the molecule, skew the electron orbits, and rotate the molecules to align with the applied field. It is for this reason that some authors prefer to call $\mathbf{D}$ the *electric displacement*.

An alternative description of the dielectric system is then given in terms of the polarisation, $\mathbf{P}$. $\boxed{\mathbf{D} = \varepsilon \times \mathbf{E} = \varepsilon_0 \mathbf{E} + \mathbf{P}}$. This polarisation model is the one used by Faraday in 1837 to explain dielectric phenomena.[4] Personally I prefer to think of the bulk material at a macroscopic level, neglecting the atomic phenomena. Thus dielectric polarisation will not be further discussed.

It is easier to think of a dielectric as being *isotropic*, meaning having the same dielectric constant in all directions. In practice, however, some materials are far from isotropic. As an example, brown rutile (titanium dioxide) has a dielectric constant parallel to its optic axis double that perpendicular to this axis. For a field of $\mathbf{E}$ which has components along both axes, the resultant $\mathbf{D}$ will not be in the same direction as $\mathbf{E}$. Thus for any anisotropic material property, the driven flux of the field will not in general be in the same direction as the driving field.

6.3 Capacitor Types

The cheapest and most common type of capacitor is the ceramic chip capacitor, typically costing as little as $0.04. Ceramic capacitors are available in three dielectric classes according to the Electronic Industries Association (EIA) standard EIA-198. Class I, ultra-stable; Class II, stable; and Class III, general purpose.

The most common Class I designation is C0G [also known as NP0]. This designation gives a dielectric spec by curve shape, not by chemical composition. Thus C0G capacitors from other manufacturers need not behave in the same way. All that can be said is that each TC curve has to fit within the +30 ppm/°C to −30 ppm/°C limits specified by the EIA.

For UHF applications there are specialist dielectric types such as porcelain and silicon dioxide. Manufacturer's data shows that the porcelain types have a significantly lower dissipation factor than ordinary C0G types, which becomes significant when designing UHF amplifiers with >1 W output.

It has been reported that some C0G formulations suffer from thermal retrace. In other words, if you heat the device up, then it cool it back down to the original temperature, the capacitance will have changed. The effect on a good quality part should be small (<0.02%), but this depends on how much ferroelectric content there is in the particular manufacturer's formulation. Barium Titanate, for example, is a very poor substance to include in a capacitor if thermal retrace is an important factor.

American Technical Ceramics have actually quoted <0.02% retrace on their porcelain capacitors, but on further enquiry this was actually a measurement resolution issue, rather than a measured effect. In any case, for C0G capacitors the amount of

[3] C.P. Smyth, *Dielectric Behavior and Structure* (McGraw-Hill, 1955).

[4] M. Faraday, 'General Results as to Induction', in *Experimental Researches in Electricity* (Taylor & Francis, 1839; repr. Dover, 1965), paragraphs 1295-1306, Vol I.

thermal hysteresis is not enough to take them outside of their ±30 ppm/°C TC band.

In practice, capacitor TCs can be specified as nominally P100, for example, which means a positive TC of 100 ppm/°C. A negative TC of 150 ppm/°C would then be N150. Hence the origin of NP0 is negative/positive zero, meaning the nominal TC is zero.

There is no specific ageing rate given for C0G capacitors. Manufacturers' data sheets state it as "none". C0G capacitors are great; they are accurate and stable. Unfortunately they are only available with values up to a maximum of 22 nF (1206 size; 100 nF, 1812). As technology moves on you will find that bigger and bigger capacitors will be available in C0G. Above this maximum it is necessary to change to a Class II dielectric.

The Class II dielectric family, called "stable", has a three letter coding scheme set by the EIA. This table applies to Class II and Class III capacitors.

1st character Low Temperature	2nd character High Temperature	3rd character Maximum Capacitance Shift
Z = +10°C	5 = 85°C	R = ±15%
Y = –30°C	6 = 105°C	U = +22% / –56%
X = –55°C	7 = 125°C	V = +22% / –82%

X7R is a common Class II spec; the capacitance stays within ±15% of its room temperature value as the temperature is varied from –55°C to +125°C. That is the only spec. There is no requirement for this change with temperature to be ***monotonic*** or for it to ***track*** from one capacitor to another. There may be some correlation between capacitors from one manufacturer, but you should not expect different manufacturer's capacitors to have the same TC slope or direction at any temperature. These capacitors must therefore never be used in situations where TC tracking is important.

Class II ceramics have a definite ageing characteristic associated with the last heat stress they experienced. When they are heated above their ***Curie point*** [perhaps somewhere in the 130°C to 170°C region] and allowed to cool, they immediately start to drift with time by a huge amount [for example –2%/(decade hour); that is –2% from 1 hr to 10 hr, –2% from 10 hr to 100 hr, –2% from 100 hr to 1000 hr &c]. Furthermore, the capacitance can reduce by up to 10% as the working voltage is increased to its maximum value. And as a final insult, the capacitance can reduce by 10% when the frequency is changed from 1 kHz to 1 MHz.

These capacitors are not stable with time, they are non-linear with voltage, their capacitance changes with frequency, and they have TCs which can only be described as horrible; this is what the EIA calls "stable"! Since the capacitance decreases with DC bias, much in the same way that the small-signal inductance of an inductor decreases with direct current bias, the small-signal distortion in a Class II/Class III capacitor can dramatically increase (say by a factor of 10×) when a DC offset is applied. In summary, Class II and Class III ceramic capacitors are not at all suitable as 'analog' components. Use them for decoupling by all means, or use them for coupling where you don't care about the coupling time-constant to an accuracy of better than about ±35%.

X7R/X5R multi-layer ceramics have replaced tantalum capacitors below 4.7 µF 16 V (1206) for decoupling purposes. The X7R/X5R ceramics are not only better on ESR and non-polar, they are also cheaper. This trend will continue to higher voltage and capacitance.

If you thought Class II ceramics were bad, then wait until you hear about Class III. The idea is to get more capacitance in a given size and voltage rating. Z5U is a popular variety; that's a temperature characteristic somewhere in the +22% to −56% region. In other words, temperature can halve the capacitance! The ageing rate is worse as well, say −7%/(decade hour), and the voltage can reduce the capacitance by 60%. Z5U types are therefore only suitable for decoupling.

Surface mount ceramic capacitors can fail open-circuit due to mechanical stress on the PCB. This might occur during assembly where break-off sections of the PCB are removed, for example. Mechanical stress can also occur due to the way the PCB is mounted into the system chassis. Realise that a surface mounted resistor or capacitor has very little ability to withstand mechanical stress through the PCB. The SOT23 and gull-wing packages used for transistors and ICs at least have small lead wires which can bend to take up any imposed strain in the PCB. However, neglecting this stress effect, the normal failure mechanism for ceramic capacitors is to go short-circuit.

Electrolytic types are used for higher capacitances (10 μF, upwards) and the vast majority of them are polarised {have + and – terminals}. If you connect them backwards they may physically explode, although large modern types have vents to allow material release in a more controlled way. It is not uncommon to see bits of capacitor flying through the air when an electrolytic capacitor is wired into a prototype the wrong way around. Since these projectiles can and do hit the ceiling, it should be clear why the use of safety glasses is considered important!

Although non-polar electrolytics have been available for decades, they have never replaced polar-electrolytics in general use. They are larger than their non-polar counterparts for the same performance and therefore more expensive for the same performance. I would strongly recommend avoiding this type of capacitor. A better solution is to use two ordinary electrolytics in series with the two negative terminals joined together, and this mid-point pulled down to a large negative rail via a 100K resistor. Each capacitor then 'sees' the correct polarity bias, despite an arbitrary DC level at the input or output to the series pair. You could instead join the capacitors by their positive leads, pulling this mid-point up to the most positive rail if this is preferable.

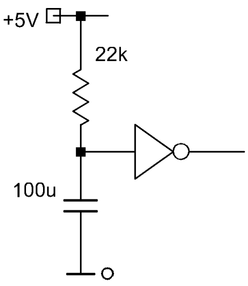
Electrolytics have significant leakage currents (microamps) which are only specified after the capacitor has had voltage applied to it for at least a minute. What this means is that if a circuit board has not been powered up for some time, perhaps months or years, the electrolytics will take significantly more than 3× their normal leakage current. If the resistor in series with the capacitor is too large, for example if somebody is trying to make a long time-delay circuit, then the voltage on the capacitor may never reach the threshold of the switching device and the overall circuit may not function at all.

Whilst mechanical systems are not normally expected to work if left unattended for some time, it is generally assumed by less technical people that an electronic system can be left on a shelf indefinitely and will work straight away when needed. This idea is both naïve and wrong.

Mechanical systems involve moving parts and it is these parts which corrode and seize-up when left unused for long periods. Electronic systems can also have corrosion problems on mechanical interconnects and switches, but these problems are generally easier to fix than for mechanical systems. Sometimes a circuit card in an edge connector may need to be removed and re-inserted so that the *wiping action* can clean the contacts. It may even be enough to just bash the side of the equipment, although this doesn't look very professional! However, electrolytic capacitors are a major source of "shelf-life" failure of previously working equipment.

High quality modern electrolytics can stand 100,000 hours shelf life at room temperature before their leakage current becomes excessive. That's around 10 years. What happens with the $0.04 capacitors you bought from the Cheapo Capacitor Corporation is somewhat less certain. Other built-in long term failures come from rechargeable cells {batteries} and lithium cells used for non-volatile retention of calibration data. In fact on-chip 'non-volatile' memory types often have a ten year retention limit as well. Thus electronic equipment should not automatically be expected to have an unlimited shelf life.

***EX 6.3.1**: This circuit has been drawn up by a digital engineer as a microprocessor reset circuit. He is about to give it to the PCB department to design into the new multi-phasic poly-fractaliser project.[‡] You have only 25 seconds to grab him before he runs off excitedly down the lab clutching this last piece of the design. What should you do? **FIGURE 6.3A:**



Any circuit which tries to produce a reliable delay greater than around one second using a single RC time-constant is doomed to failure. The fundamental problem is that the capacitor voltage will 'sit' near the switching threshold for too long and will therefore be susceptible to any microscopic amount of noise from any source. I don't care if you have a poly-brand-new-ethene [†] one farad capacitor and a one femto-amp bias current opamp; the design is fundamentally unsound for a professional design. One way to solve the long delay problem is to use a timer chip which uses hundreds or thousands of cycles of an RC discharge to produce the long delay. Such chips have been around since at least 1980 at minimal cost, and their timing is very resilient to electromagnetic interference. Another relatively inexpensive solution is to use a simple single-chip micro-controller.

Electrolytics are an important part of a design relating to power supplies for two reasons: One, they fail prematurely if you use an inadequate rating. Two, they cost a lot of money. Whilst small-signal resistors cost $0.01 and small-signal capacitors cost $0.04, an electrolytic can cost $0.20 for a small local decoupling component. Depending on the type of equipment, the cost could be $100's, but ordinarily they run to perhaps $1.

Electrolytic capacitor specs include:

- capacitance
- operating voltage
- impedance (quoted at some fixed frequency like 100 kHz)
- ESR (Equivalent Series Resistance)
- ESL (Equivalent Series Inductance)

[‡] Invented device and project.

[†] Invented solid dielectric name; brand new.

- dissipation factor, Q, loss angle
- ripple current
- ambient temperature limit
- life expectancy (could be 1000 hours at 85°C)

Because electrolytics are used for power supply decoupling and energy reservoirs, problems arise when there is too little capacitance. Thus whilst the "actual spec" is ±40%, by changing the quoted nominal value the capacitor can be specified as being −20%, +100%.

Ripple current is the RMS current flowing into the capacitor. If the ripple current is directly measured with a ***true RMS*** meter having adequate bandwidth, it is necessary to check that the meter has adequate ***crest factor*** and that the impedance of the meter does not itself reduce the ripple current.

The bandwidth and crest factor requirements are best checked by measuring the current waveform using a current probe and a scope. But since you have to use the current probe/scope combination anyway, it may be more convenient to also use them to measure the ripple current.

Failure to establish the operating ripple current could mean that the capacitor is being over-stressed without your knowledge. Unlike most other components, an electrolytic has a very definite service life. High core {internal} temperatures make them fail much faster than is acceptable. These high core temperatures are due to a combination of the operating ambient, the RMS ripple current, and the frequency of the ripple current.

If the application demands the capacitor to be discharged rapidly down to almost nothing, as occurs in photoflash applications for example, a specially designed capacitor will be needed. This sort of discharge cycle will burn out ordinary electrolytics in an unreasonably short time.

In looking at data sheets you will quickly see that a specified life of 1000 hours for an electrolytic capacitor is not unusual. 1000 hours is only 42 days of continuous use! The subject of *derating* of electrolytic capacitors is therefore of key significance to reliability, internal temperature being the key factor. If you run an electrolytic at its specified limit it is almost guaranteed to fail within a one year guarantee period. The 'trick' is to use 105°C rated capacitors and run them at less than 70°C. Always use 105°C rated capacitors for long life applications.

One manufacturer states in their data sheets[5] that the working voltage has a negligible effect on the reliability, provided that it stays within the maximum limit; the effect is apparently small compared to the limits imposed by temperature.

When you see an equivalent circuit for an electrolytic capacitor, be aware of the fact that the component *values* vary with frequency. Consider the ESR figure, *Equivalent Series Resistance*. It is not a simple resistance in series with a lossless capacitor. It is the *equivalent* resistance that could be considered to be in series with a lossless capacitor over a limited range of frequencies.

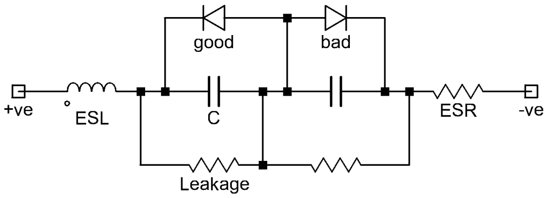
There will certainly be a fixed contribution to ESR from lead resistance. However, there will also be contributions due to dielectric loss and ***skin effect*** resistance. The net result is that the ESR is not constant with frequency. In electrolytics the ESR can decrease

[5] ELNA 1996 catalogue.

with frequency for so-called low impedance types. For example, Elna RJ3 330 µF 16 V low impedance capacitors have a quoted ESR of 0.81 Ω at 120 Hz and a quoted impedance of 0.39 Ω at 100 kHz; clearly the ESR has dropped with frequency. These capacitors are optimised for switched-mode power supplies, which is why they are specified at 100 kHz.

The rated ripple current of an electrolytic capacitor is strongly related to the frequency of the current. According to one manufacturer's data, you are allowed twice the ripple current at 100 kHz compared to that allowed at 120 Hz. This increase in allowed ripple current ties-in with the reduction in ESR. For the same internal power dissipation, double the ripple current suggests that the ESR has dropped by a factor of 4. This being the case, you have to be careful to see at what frequency the ripple current has been quoted. One manufacturer's capacitor may look better than another, simply because the ripple current has been quoted at a higher operating frequency.

FIGURE 6.3B:



The two 'plates' in an aluminium electrolytic capacitor are thin (50 µm) aluminium *foils*. The foil is etched and roughened, increasing its surface area by a large factor (100×), thereby increasing the capacitance by this same factor.

It has been known since the late 1850's that aluminium could be used as a rectifier when used in conjunction with an electrolyte. The equivalent circuit (Fig 6.3B) shows two such rectifying junctions back to back. The positive foil, the *anode foil*, is the key capacitance in the system. The electrolytic junction is *formed* by the application of just above what will be the working voltage of the capacitor and an oxide layer (12 nm/V) is grown on that foil. The negative foil is not formed, leaving that rectifier with only ≈1.5 V reverse breakdown capability.

The resulting component is not suitable for operation with reversed polarity. This fact can be demonstrated by measuring the forward and reverse leakage currents with up to ±1 V DC. The measured characteristic is very much like a lossy diode. The key thing to remember is that if the electrolytic is subjected to reverse voltage above a volt or so, the current drawn can be so great as to explode the capacitor.

Whilst it is true that an old unused capacitor can be *re-formed* by the application of the working voltage for up to an hour, this process has to be done with some current limiting device or resistor in circuit. For unused equipment, if the power supply capacitors need to be re-formed, the consequence would be that the switch-on surge current could be so excessive that the capacitors just explode, or at least fail open-circuit. Knowing this, equipment designed for decades of storage could adopt a sequenced start procedure to re-form the capacitors first before requiring normal operation.

Changing from wet electrolytics to solid film or foil capacitors, changes the spec sheets dramatically. There is no worry about self-heating, there is no ripple current, peak current or any sort of current rating. Instead there is the dV/dt rating [given as dU/dt by some manufacturers], the rate of change of voltage with respect to time.

It is often switched-mode supplies that cause solid dielectric capacitors to reach the limits of their specs and sometimes to exceed them. It is usual for switched-mode supplies to generate very fast transient voltages and the ***snubber*** circuits used to damp down the edges therefore experience very high dV/dt pulses. This is where the film/foil constructional method is used.

When making a capacitor from say polyester, you have two choices. The first way is to take a thin polyester film (<5 μm) and sputter metal onto it [molten metal sprayed on in a vacuum]; this gives a 'metallised film' capacitor. The second way is to take pieces of metal foil and wrap them up between layers of dielectric, a 'film/foil' capacitor. The names are similar and this can be confusing.

The film/foil type is larger and more expensive than the film type, but it has a superior pulse handling capability (10× better) because foil is a better conductor than sputtered film. Film/foil is also far superior to metallised film as regards harmonic distortion performance (up to 40 dB better), although manufacturers do not quote figures for this distortion. For reliable operation beyond –90 dBc harmonic distortion, do not use metallised film capacitors.[6]

dV/dt ratings of capacitors are not always openly on display; you may have to consult the manufacturer to find out what the dV/dt rating of a particular part is. This is a great trap for the unwary. If the spec was given, you might check to see how much was being applied in your application. When it is not given, it is easy to overlook. Remember, if you exceed the dV/dt rating of a capacitor it is likely to fail open-circuit or to burn out. This point is most important in switched-mode power supplies, where the dV/dt is being applied continually.

6.4 Trimmers

Capacitive trimmers are not available in values above a few hundred picofarads. Ordinarily you would not use a trim cap of more than a few tens of picofarads. Because the amount of capacitance being trimmed is so small, you will find that the action of inserting an adjustment tool into the trimmer will also change the capacitance. Thus the circuit is only 'spot on' when the trimming tool is in position. It then becomes a matter of skill to allow for the amount of shift that removing the trimming tool will make!

The circuit design should be arranged to minimise this trimming tool insertion effect. Firstly, the trimmer is often connected from some circuit node to ground. It is then sensible to make sure that the grounded end of the trimmer is the one the trimming tool touches. This precaution will minimise the capacitance shift caused by removing the trimming tool.

Using a screwdriver to adjust a trimmer is not a good idea because of all the extra metal connected to the trimmer. Thus special trimming tools are available which have a small metal insert in an otherwise plastic body. This plastic construction minimises the tool-removal capacitance shift. The metal insert is needed because plastic is not strong enough at the trimming tip. Ceramic tools are available, albeit at 10× the cost, and these are strong enough. However, even ceramic trimming tools can cause shift in the trimmer capacitance because of their high dielectric constants and also because of the pressure applied when the tool is used.

If you need to trim larger amounts of capacitance, say >100 pF, then one option is to fit extra capacitors to the PCB and link them into or out of circuit during board test. This can be done using solder links, or wire links that are optionally cut. Another choice is a

[6] C. Bateman, 'Capacitor Sounds 4', in *Electronics World*, Nov 2002.

DIP switch to ‘dial up’ more capacitance. Alternatively one can “select on test”; this could be as simple as using a capacitance box to find the correct value then soldering in the nearest fixed component available, although it has to be said that the lead-wires to the capacitance box will almost certainly cause problems on circuits with bandwidths above a few megahertz, or for capacitance values less than a few nanofarads.

Having told you how to solve the capacitance trim problem, I would strongly advise you to find another way if possible. The techniques described above are very labour intensive and therefore costly. A variable gain amplifier in the capacitor path can be used to electronically change the effective capacitance, for example.

6.5 Size matters

Roughly speaking, if you double the capacitance of a solid dielectric capacitor then its volume will double [for a given technology]. However, when you double the thickness of the dielectric to increase the working voltage, you also halve the capacitance. Hence if you double the dielectric thickness at constant capacitance the size will theoretically increase by a factor of ×4. This $volume \propto C \cdot V^2$ is a reasonable approximation for solid dielectric capacitors.

However, when an insulator gets thicker, the breakdown field strength (V/m) decreases approximately as the inverse square root of thickness;[7] thus doubling the insulator thickness may only increase the breakdown voltage by √2. This non-proportional increase in the breakdown voltage of an insulator has been known about since at least as early as 1938.[8] In practice this effect does not seem to apply to the thickness of insulator found in capacitors rated up to 2000 V; their volumes do not increase faster than the square of the voltage.

The $volume \propto C \cdot V^2$ rule does not apply to electrolytics; they follow a $volume \propto C \cdot V$ rule. As the stored energy in a capacitor is $\frac{1}{2} C \cdot V^2$ it should be clear that electrolytics store more energy per unit volume as the operating voltage is increased. This rule works well in practice up to at least 500 V.

In older text books, say prior to 1960, you will see capacitor values such as 22 mF. In those days they did not mean milli-farad, they meant micro-farad. The problem was that the mu symbol, μ, was not as easy to include in documents as it is today. Also, in those days it was rare to get such large values of capacitance. If anyone genuinely made a 22 milli-farad capacitor they would shout about it by calling it a 22,000 micro-farad capacitor. Even in the 21st century, manufacturers sometimes use *m* for micro. I therefore strongly advise against using mF for milli-farad. Use thousands of microfarads instead—it is safer.

It has been usual to use uF for micro-farad for many years. This is still acceptable, but the correct μ symbol is highly preferred. Notice that although you may not have easy access to the μ symbol, it may be available through a key sequence such as ALT+0181.

Modern capacitors now range from sub-picofarad values up to thousands of farads.

[7] 'Properties of Materials', in *Reference Data for Engineers: Radio, Electronics, Computer & Communications*, 8th edn (SAMS, 1993), p 4:17.

[8] 'Dielectric Strength' in *BR229: Admiralty Handbook of Wireless Telegraphy* (His Majesty's Stationery Office, 1938), p. 175.

Even 1 farad of capacitance was unheard of in 1980. Now you can get a 5000 F 2.5 V capacitor as a stock item from a distributor (EPCOS UltraCap; $194). The applications of these super-capacitors/ultra-capacitors include memory backup and rapid discharge situations where batteries might previously have been used. As the costs come down, the number of applications goes up.

There are two key measures to show which technologies are best for any particular application. These are energy stored per unit volume, and energy stored per unit weight. Obviously airborne and hand-carried applications require high energy density per unit weight. Most other applications stick with high energy density per unit volume.

Any table of energy / volume densities is inaccurate since manufacturers do not rate and test their batteries {cells} in the same way as each other. Also the nature of the load affects the relative performance of the various technologies. Having said that, some data is better than no data. Note that the volume energy density of NiMH rechargeable cells increased by ≈7× between 2002 and 2006.

Technology	Description	Energy (J)	Volume (cm^3)	Weight (g)	J/g	J/cm^3
Ceramic N4700	Murata DHS 8 nF 10kV	0.4	34	80	0.005	0.012
Metallised Polyester	EVOX-RIFA MDK 100 V 10 μF	0.05	3.2	4.6	0.011	0.016
Ceramic X7R	Murata GRM55 1.5 μF 50 V	0.0019	0.051	0.29	0.007	0.037
Aluminium Electrolytic	BHC-Aerovox ALS30 100,000 μF 10 V	5	107	140	0.036	0.047
Aluminium Electrolytic	BHC-Aerovox ALS30 4700 μF 500 V	587	1020	1450	0.40	0.58
Ultracap	EPCOS 2700 F 2.3 V	7140	620	725	9.9	12
NiMH rechargeable	Energizer 1.2 V 2.5 Ah AA-cell	10,800	7.4	30	360	1460
Alkaline primary	Duracell Ultra D-cell MX1300	61,000	56	145	421	1090
Lithium Primary	Energizer AA L91 Li/FeS_2	10,800	7.1	14.5	740	1500

6.6 Class X

When working on ***mains*** input suppression for a piece of equipment, there are very stringent legal requirements. Such components are safety related and therefore you need to check the current ***Regulations*** as well as understanding this text. There is a class of capacitor that is specifically designed to connect directly across the mains inputs. This is known as the Class X type [“Class-Ex”, not a roman ten.] They are much bigger than other capacitors of the same capacitance and voltage rating because they are designed to withstand severe over-voltage transients.

When you consult texts on lightning you find that it is not a question of *if* a power line will get struck by lightning, but how often it will happen, and how far away the strike will be from a particular user. There are charts available of peak transients that appear on the mains, and the rate at which they occur. As the value of the peak is increased, the probability of its occurrence goes down.

The safety standards therefore require that ***basic insulation*** on components designed to run on 230 V mains supplies be able to withstand at least 1350 V AC for 1 minute or a narrow [50 μs] transient at 2500 V. These figures may be revised [upwards] by the standards agencies to gain improved reliability. It is highly unlikely that they will ever be reduced.

An important feature of capacitors with high transient voltage-withstand capability is *self-healing*. When such a capacitor is subjected to an over-voltage event, the dielectric breaks down in a small area as a short-circuit. A suitable capacitor will then ‘clear the fault’; it will *heal* itself. Because the conducting element is relatively thin, the short-circuit causes a massive surge current which vaporises the conductor in that area. The fault is cleared at the expense of a slight decrease in capacitance. Metallised film capacitors can be very good at self-healing, but the capacitance decreases steadily with each successive transient.

Wet Aluminium electrolytic capacitors will also self-heal, but in a less destructive manner. If a weakness forms in the Aluminium Oxide insulating layer, the liquid electrolyte can move into the gap and the oxide layer will be reformed.

Whilst lightning strikes give the biggest transients, they are not the dominant source of mains transients. Failures in the power distribution network, such as flashovers and fuse breakings, can create 60 μs 1.2 kV transients. Even switching heavy plant {equipment such as motors} on or off can generate 200 μs 800 V transients.[9] These transients are clearly a problem for power supply designers, and manufacturers of inlet filters and filter capacitors. Over-voltage transients mean you cannot take components rated at $\sqrt{2}$ times the maximum RMS mains voltage and have any expectation that they will survive.

This $\sqrt{2}$ factor was commonly used on some UK domestic TVs prior to 1980, the result being that key components in the switched-mode supplies frequently failed. The peak voltage of a 230 V RMS sinusoid is 360 V [allowing for ±10% tolerance on the supply]. However, the peak rating of components for use on this supply needs to be anything from 2000 V to 6000 V to have a good chance of surviving. Fortunately, the use of a good power inlet filter reduces this peak handling requirement, the higher voltage transients being inherently narrower.

Thus Class X capacitors are so big because they are rated at much higher voltages than ordinary capacitors of the same stated working voltage. There are also subclasses of Class X capacitors. The common type used for equipment connected to ordinary wall outlets is X2. Class X1 has more stringent requirements because it is designed for connection electrically closer to the power network. A 3-phase industrial power feed would be an example of their application environment.

Now Class X capacitors are connected between the live and neutral connections. In mains filter circuits it is usual to also have capacitors to the safety ground {earth; protective conductor}. Ground is a key safety terminal for electrical equipment and it is generally designed to take at least double the rated fuse current of the equipment; the

[9] *Capacitors for RFI Suppression of the AC Line: Basic Facts*, 4th edn (EVOX-RIFA, 1996).

fuse blows well before the ground path burns out.

Continuous currents in the ground {earth} conductor of more than 3.5 mA are not tolerated without special precautions. Therefore the capacitors which connect from the mains live/neutral connections to ground have to be even safer than those directly across the mains. After all, the short-circuit failure of a capacitor directly across the mains is just a nuisance, requiring the equipment to be repaired. The short-circuit failure of a capacitor connected from live to ground is much more hazardous. Somebody might be touching the metal case of the equipment when the fault occurs.

I have already said that the ground {earth} lead is quite capable of sinking this fault current. The point is that it is supposed to be able to sink the fault current, but the fault current should not be likely to happen. This is the sort of thinking you have to adopt for safety related issues. Imagine the improbable and plan a defence against it happening. These live-to-ground capacitors are called Class Y capacitors and are clearly larger than the Class X capacitors of similar capacitance.

Again there are subclasses of Class Y; Y2 being the most common. If a Y2 fails then the current should be safely shunted to ground {earth}. This is a *basic insulation* requirement. It gives functionality rather than safety. There is little danger of electric shock because the fault current flows to ground. If the fault current could flow through a person then that requires either *reinforced* or *double insulation*. In this case a Class Y1 capacitor is required.

EX 6.6.1: A 100 nF polyester capacitor has a dV/dt rating of 25 V/µs. What is the maximum allowable transient current?

Don't think that you know enough about this subject from the brief introduction given above. If you are actually dealing with Class X and Class Y capacitors and safety related insulation, it is vital that you read the appropriate safety standards. IEC60950 is a good place to start, but there may be a better standard for your application.

6.7 Resistive Imperfections

There is no such thing as a perfect capacitor. They all have imperfections, and even the best dielectrics can have sufficiently poor characteristics that compensation or correction is necessary for measurement applications.

It is usual to see very simple models for capacitors. These generally concentrate on one or two specific characteristics of the capacitor at any one time, neglecting the rest. Looking through data sheets and text books you will see the following terms relating to the power loss in a capacitor:

- ESR
- Dissipation factor
- Power factor
- Loss angle
- Tangent of loss angle
- tan δ
- Insulation resistance
- Q
- Complex dielectric constant

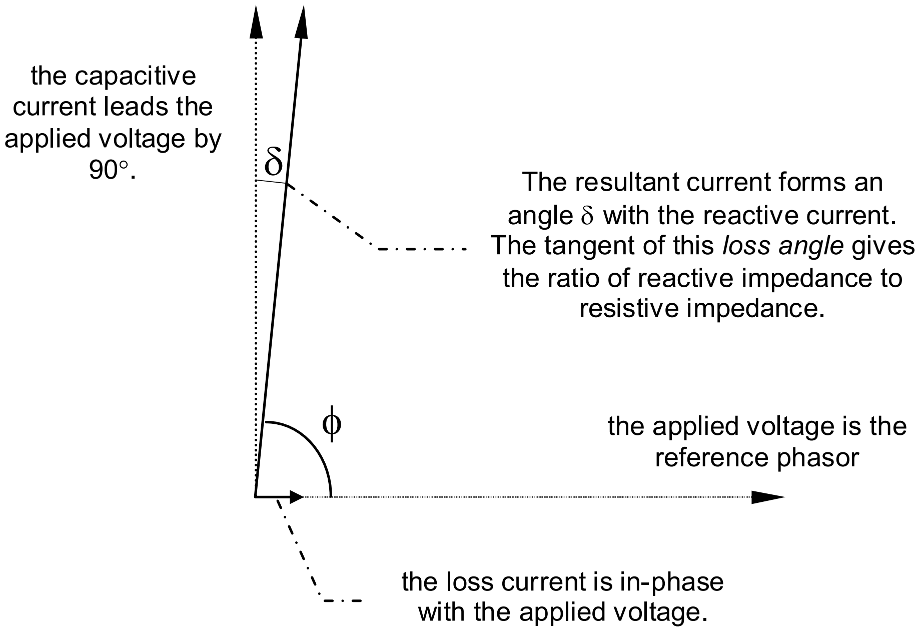
All of these relate to the resistive loss in the capacitor, since power loss in the circuit model is only through a resistive component. The ESR parameter is most often specified for electrolytics, the dissipation factor is usually given for metallised film and film/foil types, and Q is given for VHF/UHF ceramics. Complex dielectric constants tend to be given in physics and chemistry texts, where phase shift between the **E** and **D** vectors results in power loss.

The loss in an electrolytic is best modelled by a series element, whereas for a solid dielectric component, the dielectric loss is fundamentally in parallel with the capacitive element. In VHF circuits, Q is a more common circuit requirement. Whilst it is always possible to convert the parallel model to a series model, at one specific frequency, the parallel model is the naturally occurring form of the dielectric power loss.

***EX 6.7.1:** A 10 nF capacitor has an ESR of 1 Ω at 1 kHz.

a) What is the equivalent parallel resistance at this frequency?
b) If the ESR is also 1 Ω at 10 kHz, what is the equivalent parallel resistance at this frequency?

FIGURE 6.7A:


the capacitive current leads the applied voltage by 90°.
δ
The resultant current forms an angle δ with the reactive current. The tangent of this *loss angle* gives the ratio of reactive impedance to resistive impedance.
ϕ
the applied voltage is the reference phasor
the loss current is in-phase with the applied voltage.

This phasor diagram is based on a parallel model. There is a pure reactive element shunted by a pure resistive element. The phasor sum of the currents in these two ideal components gives the resultant current in the capacitor. The power loss in the capacitor can now be represented by its ***power factor***.

If you remember your first year courses, the power factor for sinusoidal circuits is $\cos(\phi)$ so that $\text{mean power} = \text{V} \cdot \text{I} \cdot \cos(\phi)$. Ideally the power factor of a capacitor would be 0, meaning no power dissipation. On the phasor diagram it is clear that ϕ, the phase angle, and δ, the loss angle, are complementary angles. Thus $\phi = 90^O - \delta$ and $\cos(\phi) = \sin(\delta)$. The power factor is therefore the sine of the loss angle.

Power factor used to be quoted for capacitors, but it is now more common to quote the dissipation factor, the tangent of the loss angle. This is the ratio of the resistive current to the reactive current, and is therefore equal to the ratio of the parallel impedances.

$$\textit{dissipation factor} = \tan(\delta) = \frac{X_C}{R_P} = \frac{1}{Q}$$

Alternatively, using a series equivalent circuit, $Q = \dfrac{X_C}{ESR}$.

The complex dielectric constant, is expressed as $\varepsilon = \varepsilon' - j\,\varepsilon''$. The imaginary part of the complex dielectric constant is 90° phase shifted relative to the real part and therefore represents a current in-phase with the applied voltage.

The more complete inter-relationship between these quantities is given by:

$$\boxed{\textit{dissipation factor} = \tan(\delta) = \frac{X_C}{R_P} = \frac{1}{Q} = \frac{ESR}{X_C} = \frac{\varepsilon''}{\varepsilon'} = \frac{\sigma}{\omega\varepsilon}}$$

The typical dissipation factor of a C0G capacitor is <0.1%, whereas it is more like 3% for an X7R. For solid dielectric capacitors the tangent of the loss angle is sufficiently low (<6%), even for lousy dielectrics, that it is reasonable to say that $\tan(\delta) \approx \sin(\delta)$ and therefore $\cos(\phi) \approx \tan(\delta)$. In other words the power factor and the dissipation factor can be considered equal for these solid dielectrics. For ordinary electrolytic capacitors, however, this approximation is less valid.

Loss Angle (degrees)	Loss Angle (radians)	Power Factor	Dissipation Factor	Q
10°	0.175	0.174	0.176 (17.6%)	5.7
5.71°	0.0997	0.0995	0.100 (10%)	10.0
5°	0.873	0.0872	0.0875 (8.75%)	11.4
1°	0.0175	0.0175	0.0175 (1.75%)	57.3
0.573°	0.0100	0.0100	0.0100 (1%)	100
0.1°	0.00175	0.00175	0.00175 (0.175%)	573
0.057°	0.00100	0.00100	0.00100 (0.1%)	1000
0.014°	0.00025	0.00025	0.00025 (0.025%)	4000

At DC the insulation resistance (leakage resistance) will give a power loss unrelated to the dissipation factor; the insulation resistance should therefore only be used at DC. Data sheets usually give insulation resistance in terms of an *insulation time constant.* Double the capacitance at the same working voltage and the insulation resistance will halve. Hence the insulation time-constant covers a whole family of capacitors having the same dielectric. This time constant might be expressed in either seconds or MΩ×μF, the two units being equal. The MΩ×μF unit is more commonly used, requiring less thought to reach the resistance value.

10,000 MΩ×μF means that a 0.1 μF capacitor has a calculated insulation resistance of 100,000 MΩ = 100 GΩ. The spec will be 'capped' by saying that the spec is the lesser of a fixed value and the MΩ×μF value. This limit stops the user from expecting unreasonably high values of insulation resistance. If insulation resistance is important for your application then make sure to look up the effect with temperature. A factor of >10 reduction over the operating temperature range is quite usual.

It is both difficult and unreasonable to make general statements that cover all dielectric types and manufacturers. However, for a given capacitor with a solid dielectric, ESR gets larger with frequency and Q gets smaller. It is useful to try various example capacitors and see what types are available. The specific types will become obsolete with time, but the tables show the variations that are possible at a given snapshot in time.

1 GHz, 10 pF, [$X_C = 15.9\,\Omega$]

Manufacturer & Type	ESR	Q
Syfer Technology: Ultra High Frequency, High-Q 0805	0.265 Ω	60
American Technical Ceramics: ATC 100A (0.055×0.055 inch) Porcelain	0.166 Ω	96
American Technical Ceramics: ATC 100B (0.110×0.110 inch) Porcelain	0.114 Ω	140
American Technical Ceramics: ATC 180R (0.070×0.105 inch) NP0 Porcelain	0.075 Ω	212

100 MHz, 100 pF, [$X_C = 5.9\,\Omega$]

Manufacturer & Type	ESR	Q
Philips Components: 0805 NP0	0.070 Ω	230
Syfer Technology: Ultra High Frequency, High-Q 0805	0.045 Ω	350
American Technical Ceramics: ATC 100B (0.110×0.110 inch) Porcelain	0.027 Ω	600
American Technical Ceramics: ATC 100A (0.055×0.055 inch) Porcelain	0.023 Ω	700
American Technical Ceramics: ATC 180R (0.070×0.105 inch) NP0 Porcelain	0.014 Ω	1135

High Q values at UHF frequencies are not easy to measure. The components are physically small surface mount devices and need to be placed in special fixtures in order for the measurement to be made. Specifically in the measurement of Q, the capacitor would ideally be parallel-resonated with a lossless inductor; the Q being the ratio $\frac{\text{resonant frequency}}{\text{bandwidth}}$, $Q = \frac{f_0}{B}$. The trouble is that inductors at UHF can be more lossy than capacitors, giving rise to a huge uncertainty in the resulting measurement. By 'huge' I mean that the result could easily be wrong by an order of magnitude.

Fortunately it has been known for many years that at UHF frequencies, transmission lines can behave as very low loss inductors. The standard method of measuring the Q of a small capacitor (<1 nF) at megahertz to gigahertz frequencies [10] is to put the capacitor

[10] J.P. Maher and others, 'High-Frequency Measurement of Q-Factors of Ceramic Chip Capacitors', in *IEEE Transactions on Components, Hybrids and Manufacturing Technology.*, CHMT-1, no. 3 (Sept 1978), pp. 257-264.

in series with an ultra low loss air-spaced coaxial transmission line. One end of the line is left open-circuit and the other end is short-circuited. The capacitor can either be put in series at the short-circuited end or in parallel at the open-circuited end. The trick now is to inject a signal and detect the resulting amplitude without loading the resonant circuit. This is done by *loose coupling* inductively {magnetically} and/or capacitively {electrostatically}. For example the voltage detection could be done with a nearby, but not touching, sensitive FET detector probe. The injection could be done by a nearby loop energised from a levelled signal generator. The Q is again determined from the ratio of the centre frequency to the 3 dB bandwidth.

This method is exemplified by the Boonton 34A resonant coaxial line and the measurement standard EIA-483. The coaxial air-line is some 3 cm in diameter and 61 cm long, looking more like a pipe than a precision measurement device. However, it is the standard method as no other technique currently gives the same accuracy on the measurement of Q at >100 MHz. Even so, the manufacturer's assessment of the uncertainty in Q is ±10% for Q=100, ±16% for Q=1000 and ±30% for Q=10,000.

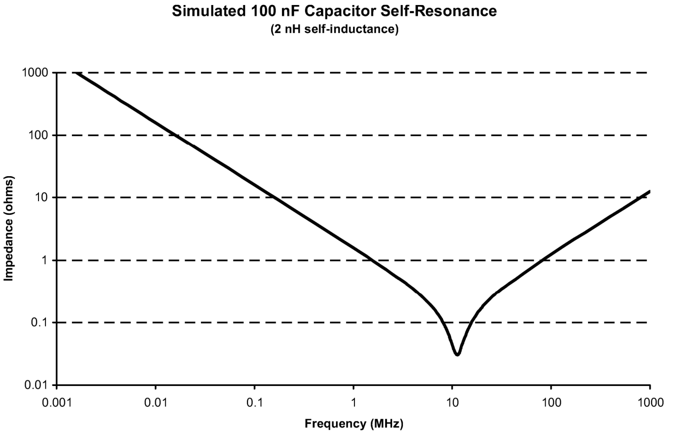
For VHF use the ESR increases approximately as the square root of frequency because it is dominated by the ***skin effect***. If the ESR were constant then the Q would drop inversely with frequency, but the increase of ESR with frequency makes the Q drop even faster.

$$ESR \propto \sqrt{f} \qquad\qquad Q \propto \frac{1}{f\sqrt{f}}$$

These formulae are excellent for ***interpolating*** manufacturers' data, but be careful when ***extrapolating*** the data. If you go higher than the manufacturers' curves you may hit the *self-resonant frequency* of the capacitor.

The self-resonant frequency (SRF; also known as the *series resonant frequency*) of a capacitor is a very important parameter. It occurs when the capacitive and inductive reactances cancel out, giving the lowest possible impedance for the capacitor. Here is a simulated curve for the impedance of a 100 nF capacitor.

FIGURE 6.7B:



If the capacitor is being used as a general purpose coupling or decoupling capacitor, operation above self-resonance does not cause a problem. The impedance starts to increase because the capacitor now behaves as an inductor, but a decoupling capacitor still attenuates, and a coupling capacitor still passes signal.

For decoupling, the phase of the attenuated signal changes by 180° around the series resonant point. It is therefore quite possible for an oscillation to be produced around the overall power system. For this reason it is not unusual to see capacitors used for decoupling that are in the 10 pF to

100 pF region for VHF capable circuitry, as well as the more usual 100 nF to 100 µF values.

Immediately above self-resonance, a capacitor is an inductor.

Some engineers deliberately use certain capacitors above self-resonance because these capacitors make good UHF sub-nanohenry inductors.[11] Capacitors as inductances have the added advantage of having a built-in DC blocking characteristic. If you keep increasing the frequency, the inductance will resonate and the capacitor will become capacitive again. However, the region well above the first self-resonance is unpredictable, uncontrolled, and unspecified; it should therefore be avoided.

If you actually measure the value of a capacitor anywhere near the self-resonant frequency you will get a strange answer. The measured capacitance increases sharply as self-resonance is approached. There is a series resonant circuit consisting of the low frequency capacitance, the ESR and the self-inductance.

$$\boxed{Z = ESR + j\omega L + {}^{1}\!/_{j\omega C}}$$

$$X_C = \frac{1}{j\omega C} + j\omega L = \frac{1-\omega L \times \omega C}{j\omega C}$$

At self-resonance the inductive and capacitive reactances cancel. This formula gives the reactance when approaching self-resonance.

This formula gives the measured *effective* capacitance below the self-resonant frequency.

$$\boxed{C_{HF} = \frac{C_{LF}}{1-\omega^2 LC} = \frac{C_{LF}}{1-\left(\dfrac{f}{SRF}\right)^2}}$$

Please be careful with the term SRF. It always stands for *self*-resonant frequency. In a capacitor the SRF and the series-resonant frequency happen to be the same. However, for an inductor the SRF is a *parallel* resonance. If you always remember that SRF stands for *self-resonant frequency* you won't go wrong.

The self-inductance of a capacitor is related to the body size and the construction; for leaded capacitors, the lead length is critically important. Thus a family of capacitors will tend to have the same self-inductance. This means that the lowest capacitance member of the family will have the highest self-resonant frequency. As data for a family of capacitors may only be given for a few particular capacitance values, this fixed self-inductance idea enables you to interpolate the self-resonant frequencies of the other family members.

Microwave circuits obviously need capacitors with exceptionally low inductance. This can be achieved using thin capacitors, 0.05×0.05×0.005 inches for example. Here the ground connection is made at the bottom, and the top surface is connected via wire-bonding or direct mounting of the microwave component. This sort of 'air-bridge' construction is expensive, but yields the necessary reduction in parasitics required for microwave operation.

To get wideband response from large capacitances it is possible to use parallel

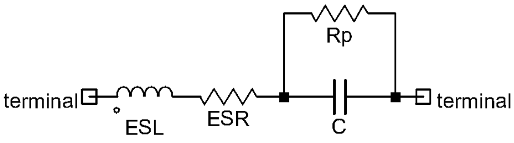
[11] V.F. Perna, 'How to Devise a Low-Loss Inductor for S-Band', in *The RF Capacitor Handbook*, rev C, 1st edn (American Technical Ceramics, 1994), pp. 6-5-2.

combinations with capacitors decades apart in value. You will see decoupling capacitors of 100 nF shunted by 100 pF capacitors to improve the high frequency response. In mm-wave circuits it is possible to get a single component with wideband response because it is internally paralleled. The 12nF//82pF buried broadband capacitor from Presidio Capacitors Inc (BB0502X7R123M16VP820) shows an insertion loss of <1 dB from 0.1 GHz to 40 GHz in a 50 Ω system. (American Technical Ceramics ATC 545L 100 nF 0402 Ultra Broadband Capacitor gives an insertion loss of <0.5 dB, typical, to 40 GHz.)

A standard capacitor larger than 10 nF will probably have a relatively low self-resonant frequency, say below 10 MHz. At these frequencies it is easy to measure the SRF by applying a signal to the capacitor from a levelled sine generator through a resistor and using a 10:1 scope probe to monitor the resulting voltage minimum. For smaller capacitors, say < 500 pF, and at higher frequencies, say >100 MHz, this method is not suitable. Some sort of test fixture needs to be made so the inductance in series with the capacitor is not increased.

One trick, used for surface mount capacitors, is to solder four identical parts end-to-end in the form of a tiny square. This ring of series-connected capacitors is then loosely coupled, magnetically, to a *dip meter*. The dip meter contains an adjustable oscillator and a voltage detector. The voltage detector responds to a nearby resonant circuit by showing a dip on a (moving coil) meter at that frequency. Knowing the low-frequency capacitance and the self-resonant frequency, the self-inductance can be readily calculated.[12]

FIGURE 6.7C:



This model is only valid over a limited range of frequencies because the components are frequency dependant. $ESR \propto \sqrt{f}$ at VHF

Decoupling capacitors reduce noise and ripple on the power rails. For optimum performance both connections to the capacitor have to be made correctly.

Always route power tracks *through* the pads of decoupling caps.

@EX 6.7.2: What does this rule mean and why is it important?

6.8 Dielectric Absorption

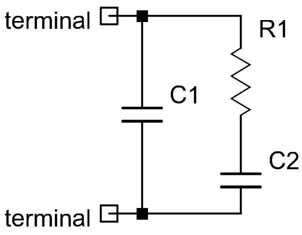
Dielectric absorption is a characteristic of any dielectric material where there is a distributed RC time-constant in parallel with the main pure capacitance. In practice then, fully charging or discharging a capacitor can take from minutes to tens of minutes, depending on the definition of the steady state condition. This effect was large [up to 20%] in historic dielectrics, but is ordinarily much less than 2%; it can even be specified as low as 0.01% for good polypropylene capacitors.

[12] R.E. Lafferty, 'Measuring the Self-Resonant Frequency of Capacitors', in *IEEE Transactions on Components, Hybrids and Manufacturing Technology*, CHMT-5, no. 4 (Dec 1982), pp. 528-530.

Faraday wrote about dielectric absorption in 1837.[13] Maxwell wrote about dielectric absorption in 1873.[14] However, some of the first relevant *quantitative* work on dielectric absorption was published in 1958 in relation to accurate capacitors for analog computers.[15]

Conduction to the inner dielectric material is very difficult because the current has to flow through a nearly perfect insulator. This ultra-high resistance causes the long time-constants that are the greatest problem with the dielectric absorption phenomenon.

FIGURE 6.8A:



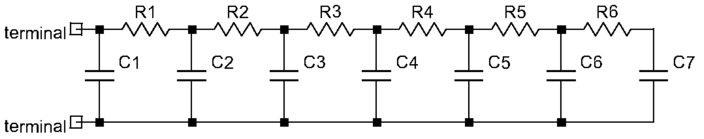
This is the simplest equivalent circuit for dielectric absorption in a capacitor.

C2 will be between 0.03×C1 and 0.0001×C1.
R1 might be greater than 100 MΩ.

In high quality film capacitors you should not expect to see a figure of more than 0.3% being quoted. Manufacturers will not publish the figure if their capacitors are poor in this respect. Dielectric absorption figures are not generally published for X7R and Z5U dielectrics, for example. One exception is American Technical Ceramics, who quote a dielectric absorption of typically 2% in their ATC200B series of capacitors. These are BX rated, the military equivalent of X7R. You should expect other X7R and Z5U ceramics to be at least as bad, if not worse, than this.

PCB materials also suffer from dielectric absorption, poor PCB material being amongst the worst capacitor dielectrics you can get. This is another good reason for drilling or milling away unnecessary PCB material around critical circuit areas.

FIGURE 6.8B:



This model is representative of the physical charge storage mechanisms.

Even 0.01% dielectric absorption, as specified on high quality polypropylene capacitors, can cause problems in high resolution (<10 ppm) sample & hold circuits, integrators and long scale DMMs. Here is the typical problem: You discharge a capacitor completely by placing a low value resistor across it for a very long time, this time being long enough for the current flow to have "ceased". Depending on your definition of the point at which "no current" is flowing, the discharge could easily take an hour or more. You then charge the capacitor quickly to an applied voltage for say for a few seconds, and then

[13] M. Faraday, *Experimental Researches in Electricity* (Taylor & Francis, 1839; repr. Dover, 1965), pp. Paragraphs 1245-1249, Vol I.

[14] J.C. Maxwell, 'Absorption of Electricity', in *A Treatise on Electricity and Magnetism*, 3rd edn (Clarendon Press, 1891; repr. Dover Publications, 1954), pp. Para 53, Vol I.

[15] P.C. Dow, 'An Analysis of Certain Errors in Electronic Differential Analysers II - Capacitor Dielectric Absorption', in *Institute of Radio Engineers: Transactions on Electronic Computers* (Mar 1958), pp. 17-22.

you leave it on its own for a while to see what happens. The voltage will change by reason of the dielectric absorption effect. The charge is redistributed to the parallel equivalent capacitors and therefore the voltage on the capacitor drops. This is a typical sample & hold application.

To measure the effect you would do the experiment the other way around. You would charge the capacitor up to a fixed value for a very long time, then discharge it to 0 V for a short time and look at the *reappearance* of voltage on the capacitor. This is a very much easier thing to do because you may be looking for only 0.01% of the original voltage as the change. It is easier to measure this small change from ground than it would be to measure it 'standing on top' of a 100V DC level.

The dielectric absorption problems associated with high-voltage (kilovolt) capacitors are somewhat different. These capacitors become dangerous due the *recovery voltage*. Consider a capacitor that has been charged up to 20,000 V. If the dielectric absorption is 3%, the recovery voltage is 600 V. Depending on the capacitance, this much voltage will be somewhere between dangerous and lethal. Thus capacitors with kilovolt ratings should be supplied with the terminals short-circuited by a conductive strap, thereby allowing them to be safely handled.

One classic risk of this recovery phenomenon is on the final anode of cathode ray tubes, particularly as used on domestic colour TV sets. It is important to always discharge the final anode down to the chassis ground immediately before removing the anode cap. The final anode runs typically runs up to 25 kV. It is better to learn this one from a book, rather than to take a shock off the tube and learn the hard way. Ground the shaft of a long thin screwdriver, using a flying lead to the chassis, and slide the screwdriver under the final anode cap until it touches the anode. You may hear a little electrical 'crack' as the anode is discharged.

Here is a method of quantifying the dielectric absorption based on MIL-PRF-19978J (clause 4.7.17).

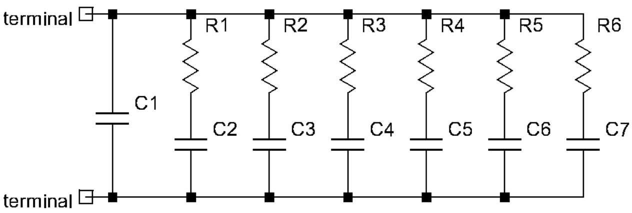
i. Charge the capacitor to its rated voltage for one hour, ensuring that the initial surge current does not exceed 50 mA.
ii. Disconnect the charging source, then discharge the capacitor though a 5 Ω resistor for 10 seconds.
iii. Disconnect the 5 Ω resistor and measure the recovery voltage with a meter having an input resistance ≥10 GΩ. Record the maximum reading over the next 15 minutes.
iv. The dielectric absorption is defined as the ratio of the maximum value of the recovery voltage to the charging voltage, expressed as a percentage.

$$\boxed{\text{dielectric absorption (\%)} = 100 \times \frac{\text{maximum recovery voltage}}{\text{charging voltage}}}$$

The good news is that the time constants involved in this phenomenon are of the order of magnitude of seconds. Thus as the cycle time reduces, the dielectric absorption effect reduces almost in direct proportion.[16]

[16] R.A. Pease, 'Understand Capacitor Soakage to Optimize Analog Systems', in *EDN*, Oct 1982.

FIGURE 6.8C:



This model uses a sequence of capacitors C2, C3 &c, which get progressively smaller, eg a factor of two between adjacent capacitors. The resistors R2, R3 &c get progressively smaller as well, since the dielectric absorption error gets less at shorter time intervals.

When loaded PCBs are ultrasonically cleaned to remove flux, the solvent and flux can penetrate the PCB material and make the dielectric absorption much worse than the basic PCB material. The trend in manufacturing is to use "no-clean" fluxes, which eliminate this cleaning step, but which may leave some slight surface contamination, leading to excess board leakage and dielectric effects.

EX 16.8.1: According to simple theory, how many time constants are required to discharge capacitance C through resistance R, starting from an initial value of 5 V and ending when the voltage reaches 100 μV? How long would you actually expect to have to wait?

6.9 Noise

Everybody knows that resistors suffer from Johnson noise. A capacitor can also be assigned a noise voltage as well. The purpose of this is to eliminate parts of the circuit that have capacitors in them from a noise analysis. Rather than looking at resistor values, it is easier to look at the value of the decoupling capacitor and then eliminate that noise source.

It should be clear that any capacitor has a leakage resistance. It may be huge, but it is not *infinite*, and even if it *tends* to infinity this argument still holds good. Consider an unspecified shunt resistor R across the capacitor C.

The Johnson noise formula is: $V_{RMS} = \sqrt{4kTR\,\Delta f}$

The 3 dB bandwidth B is preferable to the Δf term, giving: $V_{RMS} = \sqrt{2\pi kTR\,B}$ for a single-pole roll-off.

The RC network has a 3 dB bandwidth of $B = \dfrac{1}{2\pi CR}$, resulting in $\boxed{V_{RMS} = \sqrt{\dfrac{kT}{C}}}$

Einstein [17] produced this formula in 1907 using the *equipartition law.*

[17] Van Der Ziel, A., 'History of Noise Research', in *Advances in Electronics and Electron Physics*, 50 (1980), pp. 351-409.

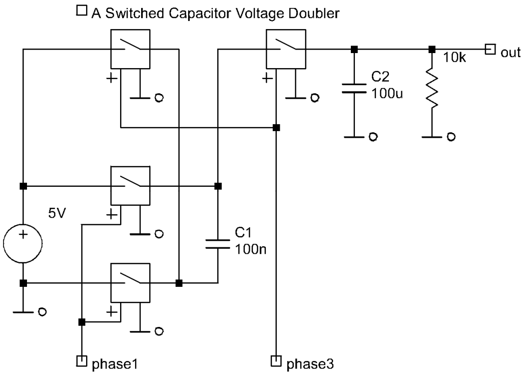
In any situation the Johnson noise voltage across a capacitor will always be less than or equal to the value given above. This table is for use at room temperature. I have conveniently neglected the possibility of excess noise in the capacitor. It is possible that a poor capacitor could have a greater amount of noise than that given above, particularly when a voltage is applied across it.

capacitor	maximum RMS noise
1 pF	64 μV
10 pF	20.4 μV
100 pF	6.4 μV
1 nF	2.0 μV
10 nF	0.64 μV

6.10 Switched Capacitors

EX 6.10.1: A simple switched-capacitor voltage doubler is used to generate an approximate 10 V rail from a 5 V rail. The output capacitor is 100 μF and the 'flying capacitor' is 100 nF.

FIGURE 6.10A:



This circuit uses a 4-phase non-overlapping clock with equal periods for the charge and discharge periods. There are dead periods on either side to prevent the switches shorting-out the power rails. If the circuit operates at 50 kHz with a 1 mA load (10K), what is the best efficiency that can be achieved (neglecting the losses in the control circuitry)? Neglect tolerances, and ESR in the capacitors.

Switched capacitors are used extensively in precision multi-pole monolithic filters. A switched capacitor can be made to 'look like' a variable resistor according to the clock speed. Thus variable coefficient RC filters can be made with accurate ratios of values. This tight matching is essential for filters with four or more poles. These filters are not ideal, however. They suffer from noise, in the form of clock breakthrough, and harmonic distortion above a few tens of kilohertz. Getting better than –70 dBc harmonics at 50 kHz is difficult.

6.11 Design Maintenance

Design maintenance relates to all components. Here is a typical problem. You produce a design, then you manufacture a product for a year or more, and you give the finished product a 5 year warranty. You are therefore obliged to maintain component availability for a minimum of 5 years and probably more like 10. Maybe you have produced a good design which runs in production for 5 or more years and then has to be maintained for another 5 years after production ends. It still amounts to at least 10 years of product maintenance.

Amongst resistors, capacitors, inductors, transformers, fuses, filters, fans, diodes and transistors, the worst of these for becoming obsolete is probably electrolytic capacitors. The reason is that the manufacturers are always improving them!

By 'improving' I mean the capacitance per unit volume at a given working voltage is increased, the ESR is reduced, the impedance is reduced, the ripple current is increased, the lifetime is increased, the operating temperature is increased, or the leakage current is reduced. If any of these good things is achieved, the manufacturer will make the old range obsolete and offer you a new range.

Now you may think that this is wonderful and that you can just put the new part in without any worries, but you would be wrong. This is where good documentation helps you out. After all, you may be picking up somebody else's design and you don't know what was required in the original design. In this case, picking the replacement can be time consuming and therefore costly. In any case, if it is 5 years since you designed something, it can be hard to dig up the design notes as to why a particular component was chosen. For this reason I like to see an archived electrolytic capacitor description that contains the following information:

- Capacitance and tolerance
- Working voltage
- Ripple current (at a specified frequency)
- Size (diameter and height)
- PCB mounting pitch
- ESR or impedance (if it is used in a low ESR or low impedance application)
- 85°C or 105°C rated part?

You will thank me when you come to need a replacement and you find the old part is not in the current catalogues, the old catalogues having been thrown out. Alternatively, archive the spec sheets on every component that you use.

For replacement purposes you should do the following: Ensure that the capacitance is the same. There is no guarantee that extra capacitance will work. For example doubling the capacitance may make the start-up current too high, or it may slow down the start-up voltage swing, making the sequence of power rail voltages go out of spec. Also more capacitance can reduce the *conduction angle* in the bridge rectifiers causing them to overheat.

Next, ensure that the ripple current rating is the same or higher than the old component. A higher ripple current rating should not cause a problem. A lower ripple current rating can obviously cause a problem. Just because you get more capacitance per unit volume with a new capacitor type, don't automatically expect the ripple current to be the same. Actually the smaller case size parts generally have a lower ripple current rating anyway, so you may need to increase the working voltage in order to get the same or better ripple current rating.

The only slight caution is that some designs of low drop-out regulators (LDO) rely on the *imperfections* in the output capacitors to remain stable. A lower ESR capacitor in this application may make the regulator unstable. There are two things to do about this:

- ☺ Don't design circuits that rely on how bad the components are because they may get better. The minimum ESR is never quoted, only the maximum.
- ☺ Try the replacement component in the circuit before you commit yourself to buying thousands of them.

Adequately documenting the parts used in a design makes subsequent design maintenance a much simpler task. It also means that people other than the original designers can more readily maintain the product; an important concept since staff may move on, retire, get run over by a bus, or simply drop-dead during the course of a project.

CH7: the inductor

7.1 Comparison

An *inductor* is an electrical component having *inductance* as its dominant electrical attribute, the term *inductance* being introduced by Heaviside in 1886 to replace the earlier cumbersome term *coefficient of self-induction*.

In the same way that the cheapest capacitor costs perhaps 4× more than the cheapest resistor, the cheapest inductor typically costs 4× more than the cheapest capacitor.

INDUCTOR	CAPACITOR
magnetic flux	electric flux
operates on current	operates on voltage
inductance is non-linear with current	capacitance is non-linear with voltage
ferromagnetic core gives more inductance per unit volume	dielectric gives more capacitance per unit volume
ferromagnetic properties drop dramatically above the *curie point*	ferroelectric properties drop dramatically above the *curie point*
voltage breakdown is a secondary consideration	current damage is a secondary consideration
bigger for low frequency operation	bigger for low frequency operation
bigger for higher current operation	bigger for higher voltage operation
cost increases with size	cost increases with size
measured value increases as self-resonance is approached	measured value increases as self-resonance is approached

In its simplest form an inductor is a piece of wire, usually coiled. Just as a capacitor uses a dielectric to achieve higher capacitance, an inductor uses a ferromagnetic core to achieve higher inductance.

The dielectric increases the *electric flux* and the ferromagnetic core increases the *magnetic flux*. There are many analogies between inductors with ferromagnetic cores and Class III ceramics as seen in this table.

Techniques for producing a low loss (high Q) inductor at 300 MHz are quite different to those used at 1 kHz.

7.2 Losses

There is no ferromagnetic core material which does not suffer from *hysteresis loss*. This is the B-H loop seen in physics lessons and first year electrical engineering classes. Every time the material is driven around the B-H loop, energy is dissipated in the core. Driving the material around the B-H loop faster therefore dissipates more energy per unit time. The hysteresis loss should therefore increase at least proportionately to frequency.

Since the magnetisation of a material consists of re-aligning *magnetic domains*, this process may become more difficult at higher frequencies. In this case the hysteresis loss will increase more rapidly than proportionately to frequency.

The other major core loss contribution is due to *eddy currents*. For a conductive core material like silicon iron, as used in mains-frequency power transformers, eddy current loss is minimised by splitting the core up into thin sheets (≈0.5 mm thick), electrically insulated from each other. These thin sheets are known as *laminations*. The lamination is done parallel to the magnetic flux, the eddy currents 'wanting' to flow in coaxial circular paths around the flux lines. Eddy current losses are proportional to both frequency squared and flux density squared.

For iron based cores, eddy current loss is a function of the construction of the core and as such it is a controllable factor. Specifically, the use of thinner laminations dramatically reduces the eddy current loss, the power loss per unit volume being roughly proportional to the lamination thickness squared. Hysteresis loss, however, is only a function of the material and the flux density. Steinmetz produced a formula for the hysteresis loss, circa 1892: $P_H = k \cdot v \cdot f \cdot \hat{B}^x$

P_H	=	power dissipated as hysteresis loss.
k	=	a constant for a particular core material, the *Steinmetz coefficient.*
v	=	the volume of the core material.
f	=	the frequency.
$\hat{B}$	=	the peak flux density in the core.
x	=	a constant for the core material, typically 1.6 for iron cores.

Since ferrites have high DC resistances, they do not need to be laminated to minimise eddy current losses. Thus it is neither necessary, nor useful, to try to break down core losses into hysteresis loss, eddy current loss and residual loss. An empirical formula for Philips 3C8 ferrite is:

$$\boxed{P = k \cdot v \cdot f^{1.3} \cdot \hat{B}^{2.5}}$$

The exponents of the frequency and flux density terms can be read off the manufacturer's curves for the material when plotted on log-log scales. For other ferrites from the same manufacturer, the spread of exponents for the $\hat{B}$ term is between 2.2 and 3.0 [applies to 3C85, 3C90, 3F3 and 3F4]. The value 2.5 is a good starting point for a rough analysis. For the frequency term, the exponent is between 1.3 and 1.5 on these same ferrites. Such a formula is only useful until the core becomes saturated, say < 200 mT.

The formula becomes useful when designing magnetic parts for switched-mode power supplies. For a given flux, ϕ, if the core is made twice as wide, the flux density halves and the volume doubles. Provided the index of peak flux density in the power loss formula is greater than 2, the core loss is more than halved by doubling the core size.

There is a slight problem with the power loss formula with respect to ***dimensional analysis***. The non-integer exponents of B and f make a mess of the dimensions of the constant k. One solution to this problem is to normalise the power to a reference frequency. The power loss is measured under a specific set of reference conditions and the actual power loss is evaluated by extrapolation. Using the subscript *R* to denote reference conditions:

$$\boxed{P = P_R \cdot \left[\frac{v}{v_R}\right] \cdot \left[\frac{f}{f_R}\right]^{1.3} \cdot \left[\frac{\hat{B}}{\hat{B}_R}\right]^{2.5}}$$

The terms in square brackets are all dimensionless ratios, eliminating the dimensional analysis problem.

VHF/UHF inductors may not have ferromagnetic cores, but they still have losses. The DC resistance of any inductor is only useful for calculating the DC volt drop across the inductor. The actual VHF+ equivalent series resistance will be due to the dielectric loss in the *coil former* {bobbin}, and the resistance of the thin surface layer (***skin effect***). The loss will be increased if the turns are tightly packed (*proximity effect*). This concentrates surface current in small areas, thereby increasing the effective resistance. One remedy is to space out both the turns and the layers.

The ultimate in lossy inductors is the ferrite bead. When used for suppression purposes, the ferrite bead is optimised for maximum RF resistance with minimal DC resistance. Ferrite beads are useful at circuit level for increasing the impedance of current sources, isolating bias components from the signal path and preventing oscillations. It is easy to get 1 kΩ of resistance at UHF frequencies with 1 Ω DC resistance. Ferrite tubes and blocks are also useful around cables to discourage unbalanced RF signals in the cables (any RF current is encouraged to both *go* and *return* down the same cable).

7.3 Self-Inductance

For a single-turn circuit, inductance is the constant of proportionality between the current and the total magnetic flux that it produces, $L = \frac{\phi}{I}$. Thus integrating the flux density, **B**, created by a current, *I*, gives the inductance.

Ampère's Circuital Law says that the closed-path line integral of magnetic field intensity, **H**, is equal to the current enclosed by that path. (*s* is distance in this integral).

$$\oint_s \mathbf{H} \cdot d\mathbf{s} = I$$

For simplicity, a circle, centred on the wire's axis, and perpendicular to the axis, is chosen as the closed path of integration. This path is always in the same direction as **H**, allowing the *dot product* to be replaced by a multiplication with the path length, $2\pi r$.

Thus $H \times 2\pi r = I$, giving $H = \frac{I}{2\pi r}$.

Looking at the cross-section of a thin coaxial cylinder of thickness δr outside of the wire, the flux enclosed is found by using $\mathbf{B} = \mu\mathbf{H}$ and the above formula for *H*:

$$\delta\phi = B \times [area] = \mu H \times \delta r \times [axial\ length] = \frac{\mu I}{2\pi r} \times \delta r \times [axial\ length]$$

$$\frac{\text{inductance}}{\text{axial length}} = \int \frac{d\phi}{I} = \frac{\mu}{2\pi} \int_r^\infty \frac{dr}{r} = \frac{\mu}{2\pi} [\ln(r)]_r^\infty = \frac{\mu}{2\pi} \ln\left(\frac{\infty}{r}\right)$$

This result suggests that the inductance per unit length of a long isolated wire is infinite! In the calculation of inductance one makes the assumption that the phase of the current is constant at all points in the conductor, and that the magnetic field is in phase with the current. When these approximations are valid the situation is referred to as *quasi-static*. The inductance is a simplification of transmission line effects, the finite velocities of propagation of the field and the current being conveniently neglected. It is therefore not reasonable to sum a quasi-static field to infinity. Eventually the approximation of "no phase shift" of the field will break down and the flux integral will not be valid. The infinite result obtained above was simply due to an incorrect mathematical model.

With a coaxial arrangement of the *go* and *return* conductors, the return current in the outer conductor only creates a magnetic field outside of itself, and this field exactly cancels the field due to the inner conductor. The magnetic field outside of the coaxial arrangement is therefore (ideally) zero and the inductance per unit length is given by the above formula, limited by the radius of the outer conductor, but with an additional correction for the flux linkage within the body of the wire.

In the integration above, the lower limit was taken as the outer radius of the inner conductor. However, if the current is uniformly distributed over the cross-section of the wire, there will be circles of flux within the conductor linked to the current they surround. This inner flux linkage is usually referred to as *internal inductance* and is accounted for in the formula below by the inclusion of the ¼ term. Since high frequency current progressively crowds towards the outside of the inner conductor (skin effect), the internal inductance reduces with increasing frequency.

$$\boxed{L_{COAX} = \frac{\mu}{2\pi}\left[\frac{1}{4} + \ln\left(\frac{R}{r}\right)\right] = 2\cdot\left[\frac{1}{4} + \ln\left(\frac{R}{r}\right)\right] \text{ nH/cm}}$$

r = outer radius of the inner conductor; R = inner radius of the outer conductor.

The inductance of a piece of straight wire is calculated by neglecting the flux past the ends of the wire.[1]

The self-inductance formula for an isolated straight piece of circular wire is:[2]
length > diameter

$$\boxed{L = 2\times\left[-\frac{3}{4} + \ln\left(\frac{4\cdot length}{diameter}\right)\right] \text{ nH/cm}}$$

Using the formula above, 1 cm of 0.5 mm diameter wire has a self-inductance of 7.3 nH. This can be a significant amount of inductance.

***EX 7.3.1:** A 50 Ω (output impedance) RF signal generator is set to give 1 V into a 50 Ω load. The output is 'shorted' to ground via 10 mm of 0.5 mm diameter wire. What will the signal be at the output of the generator at:

a) 1 MHz ?
b) 100 MHz ?
c) 1 GHz ?

For rectangular conductors the inductance formula is very complicated, but it can be greatly simplified provided accuracy of ±0.3% is acceptable:

$$\boxed{L = 2\times\left[\frac{1}{2} + \ln\left(\frac{1.998\times length}{width\times depth}\right)\right] \text{ nH/cm}}$$

For any shape of conductor, additional paths of current flowing in the same direction as the original piece assist the flux. Their *mutual inductance* adds to the self-inductance. However, when the conductor is looped back on itself to form a complete circuit, the mutual inductance *subtracts* from the self-inductance.

It is important to realise that a "single isolated conductor" is a meaningless situation. Always consider a complete circuit. It is easy to do this by considering a long parallel pair of circular conductors. At high frequencies, the skin effect means there is little current in the middle of each wire, so the internal self-inductance becomes small enough to neglect. The term 'high frequency' here is evaluated by comparing the skin depth to the wire radius. When the skin depth is several times smaller than the wire radius, the

[1] A. Gray, 'Calculation of Inductances', in *Absolute Measurements in Electricity and Magnetism*, 2nd edn (Macmillan and Company, 1921; repr. Dover Publications, 1967), pp. 493-495.

[2] R.M. Wilmotte, 'Self-Inductance of Straight Wires', in *Experimental Wireless & The Wireless Engineer*, 4 (June 1927), pp. 355-358.

internal self-inductance can be neglected. The high frequency inductance per unit length in this case is given as:

$$\boxed{L = 4 \times \ln\left(\frac{2s}{d}\right) \text{ nH/cm}}$$

where s is the axial separation of the wires, each of diameter d. This formula is derived on the assumption that the wires are sufficiently far apart that the current distribution within the wire is radially symmetric. More exact analysis takes into account the re-distribution of current density which occurs when the wires get close together. Such redistribution of current is known as the *proximity effect.* A more exact formula for the high frequency inductance of parallel wires is given by:

$$\boxed{L = 4 \times \ln\left(\frac{s}{d} + \sqrt{\left(\frac{s}{d}\right)^2 - 1}\right) \text{ nH/cm}} \equiv \boxed{L = 4 \times \operatorname{arccosh}\left(\frac{s}{d}\right) \text{ nH/cm}}$$

One additional centimetre of 1 mm diameter wire, separated from the return wire by 1 cm, gives an additional inductance of 12 nH. This parallel wire formula gives a physically real meaning to the inductance of an additional length of wire put into an existing circuit. A useful approximation for inductance of any piece of wire is therefore roughly 10 nH/cm (1 nH/mm).

7.4 Resonance & Q

The *loss* element in an inductor can be represented by either a series resistance or a parallel resistance, but the value of these equivalent resistances will change with frequency. At the mains {power line} frequency, the power loss in an inductor for a given current is important, making a series model the most appropriate.

For small-signal applications, the ratio of reactive to resistive impedance becomes important; Q is then the relevant figure of merit.

$$\boxed{Q = \frac{\omega_o L}{R_S} = \frac{2\pi f_0 L}{R_S}}$$

Assuming the inductance is constant with frequency, the Q will increase with frequency provided R does not increase at the same rate. If the resistance is primarily due to the ***skin effect***, the resistance should not increase faster than the square root of frequency. In this case the Q should rise with the square root of frequency. This characteristic can be seen on real VHF inductors.[3]

[3] EXAMPLE: AVX Accu-L 0805 inductors. 1.2 nH to 22 nH in the range 150 MHz to 1 GHz

In any case, the Q of an inductor increases with frequency up to a maximum value then falls off rapidly to zero at its *self-resonant frequency* (SRF). Above self-resonance the inductor appears to be a capacitor. Note that at self-resonance an inductor has its maximum impedance.

FIGURE 7.4A:

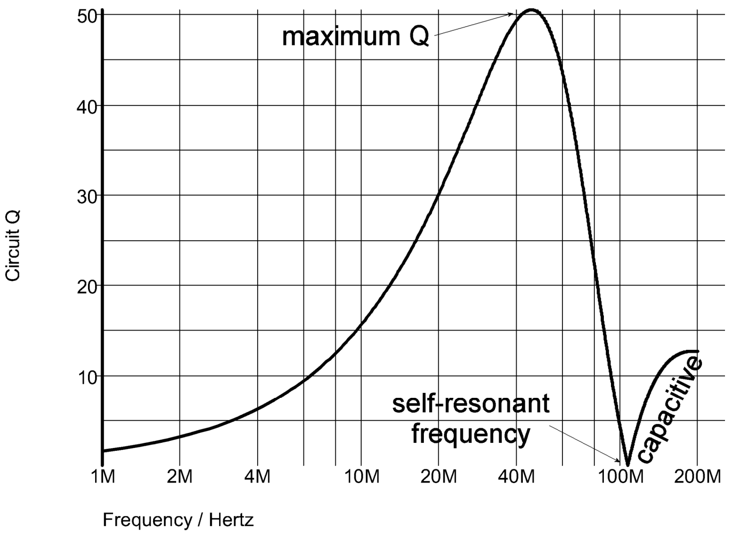
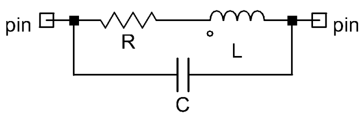
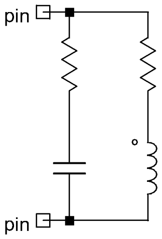


FIGURE 7.4B:



This is the simple inductor model given in text books and data sheets. The distributed capacitance within the inductor has been modelled by a single lumped capacitance.[4] Unfortunately this model predicts a higher than normal Q above the self-resonant frequency. It also does not simulate the relative positions of the Q-peak and the self-resonant frequency.

FIGURE 7.4C:



This slightly better model uses an extra resistor to damp the Q above self-resonance. The Q-peak and self-resonance of the circuit can then be separated by as much as 5:1.

EX 7.4.1: What is the approximate equation for the self-resonant frequency in terms of the equivalent circuit elements for a high Q inductor,

a) using the 3 element model?
b) using the 4 element model?

EX 7.4.2: [special interest only] Derive an equation for the relationship between the Q-peak and the self-resonant frequency for the three element model.

The relationship between the resistors and the Q-peak on the 4-element model is not mathematically straightforward. Also the 4-element model does not fit the Q-curve satisfactorily. The previous exercise showed that the Q-peak for the 3-element inductor model is such that $\frac{SRF}{f_{\hat{Q}}} = \sqrt{3}$. In practice, even for air-cored inductors, the Q-peak can occur at a factor of 2 to 9 lower than the SRF. For shielded iron or ferrite inductors, this factor can get at least as large as 50. Thus a more comprehensive model is needed.

[4] F.K. Kolster, 'The Effects of Distributed Capacity of Coils Used in Radio-Telegraphic Circuits', in *Proceedings of the Institute of Radio Engineers*, 1 (1913), pp. 19-34.

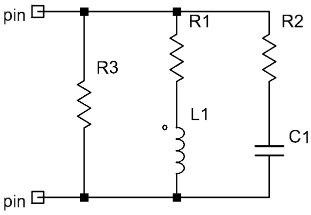
$$\boxed{L = \frac{L_{LF}}{1-\left(\frac{f}{SRF}\right)^2}}$$

Measured inductance increases as the self-resonant frequency is approached. For less than 1% error the inductance has to be measured at a frequency lower than 10% of the self-resonant frequency. This formula clarifies what is meant by 'low frequency inductance'.

EX 7.4.3: [special interest only] Derive the above equation for the measured inductance of an inductor. Hint: assume the Q is infinite.

A better model of an inductor is achieved using an additional element.[5]

FIGURE 7.4D:



The series resistor, R1, causes the Q to increase with frequency. The parallel resistor, R3, causes the Q to decrease with frequency. Having both allows the Q maximum to be placed much lower down than $SRF/5$.

If the maximum Q occurs at $SRF/10$ or lower, it is possible to use very simple formulae to calculate the values of R1 and R3.

$$\hat{Q} = \frac{1}{2}\sqrt{\frac{R_3}{R_1}}\ ;\ f_{\hat{Q}} = \frac{\sqrt{R_1 R_3}}{2\pi L}\ \text{giving}\ R_1 = \frac{2\pi f_{\hat{Q}} L}{2\hat{Q}}\ \text{and}\ R_3 = 2\hat{Q} \times 2\pi f_{\hat{Q}} L\,.$$

EX 7.4.4: [special interest only] Derive the above relationships, neglecting self-resonance entirely. Note that the little 'hat' over the Q means 'the maximum value of'.

The useful frequency range of an inductor is governed by the application. Suppose the inductor introduces bias to a circuit. In this case only the impedance magnitude is important. The inductor can then be used up to *and beyond* its self-resonant frequency.

Go too far beyond the SRF and the inductor will hit the first series resonance; this may not be a disaster, however, as the minimum impedance is not usually very low. Even without impedance curves, up to 1.5× the minimum quoted SRF is fairly safe. It is correct to estimate that the impedance is larger than $2\pi f L$.

For use in a tuned circuit, the inductor has to be used below its self-resonant frequency. The SRF on an inductor is not given with any accuracy by manufacturers. It is always given as greater than some value. On the other hand, the inductance may be given to ±2%. Thus a resonant circuit can only be toleranced when using the inductor well below self-resonance.

The capacitance internal to the inductor has a poor Q, as evidenced by the large resistor in series with it in the model. Using an external high-Q capacitor increases the Q of the combination *beyond* the Q of the inductor alone. The optimum resonant frequency of the pair is some small amount higher than the Q-peak frequency of the inductor alone.

From the manufacturer's point of view, they want to write the highest Q figure they can. This suggests that they will use a high test frequency, close to the Q-peak point. The

[5] L.O. Green, 'RF-Inductor Modeling for the 21st Century', in *EDN* (Cahners), Sep 27, 2001, pp. 67-74.

problem is that they quote the same test frequency for a range of inductor values. Thus you can't be sure how close to the Q-peak they are actually measuring the particular inductor of interest. Also, there may be a limitation in their test equipment. The things to be on the lookout for are a high test frequency as a ratio of the SRF, a high SRF, and a good minimum Q spec. Ideally the test frequency will be close to the frequency at which you wish to use the inductor. This then gives less uncertainty for your application.

The use of an actual inductor in a resonant circuit at UHF is not going to give the best possible results. You can get a better inductor using a length of transmission line if you can afford the space.[6] In fact a transmission line could replace the whole resonant circuit. As a general rule, you can get a better Q by using up more space. An 0603 sized UHF inductor usually gives a worse Q than an 0805 for example. As frequencies head up into the microwave and mm-wave region the highest Q circuits are achieved by using cavities.

7.5 Air-Cored Inductors

Air-cored inductors are useful for tuned circuits, not least of which is because of their inherent linearity and therefore their lack of ***intermodulation*** products. The use of an air core also means no core losses and therefore the possibility of higher Q above a few tens of megahertz.

The formula for the low-frequency inductance of a single-layer cylindrical coil (***solenoid***) is: $\boxed{L = F \cdot n^2 d \text{ nH}}$. The inductance in nH is equal to the product of the coil diameter, *d*, in mm; the turns ratio squared; and an aspect ratio factor, *F*, based on the ratio of the coil's diameter to its length. The inductance is proportional to the diameter of the coil, if the number of turns and shape stay the same.

$\frac{\text{diameter}}{\text{length}}$	*F* nH/mm
0.1	0.09463
0.5	0.4037
1	0.6794
2	1.037
2.5	1.164
10	2.007

The aspect ratio factor *F* is ultimately derived from work done by Lorenz (1879) and a table of values published by Nagaoka (1909). The inductance calculation is based on a circular solenoidal current sheet. The formula contains ***elliptic integrals*** and is therefore not in a convenient form for general use. Tables have been essential in the past to minimise the calculations required. Nagaoka's table has been republished in a large collection of inductance tables and formula.[7]

Wheeler gave a computationally simple approximation to Nagaoka's formula,[8] applicable to coils with a diameter/length ratio of less than 2.5.

Using more optimised coefficients, Wheeler's formula becomes:

$$\boxed{F = \frac{0.987}{0.45 + \frac{\mathit{length}}{\mathit{diameter}}} \text{ nH/mm}}$$

$$\frac{\mathit{diameter}}{\mathit{length}} < 2.5$$

[6] F.E. Terman, 'Resonant Lines in Radio Circuits', in *Electrical Engineering*, LIII (July 1934), pp. 1046-1053.

[7] F.W. Grover, *Inductance Calculations: Working Formulas and Tables* (1945; repr., Imp. Hai-phong, 1970).

[8] H.A. Wheeler, 'Simple Inductance Formulas for Radio Coils', in *Proceedings of the Institute of Radio Engineers*, 16, no. 10 (Oct 1928), pp. 1398-1400.

The above formula is accurate to ±0.3% *relative to Nagaoka's formula*. The accuracy figure is not an absolute amount, as it does not take into consideration the problems with the original assumptions and limitations. Lundin's formulae (1984) are accurate to better than 30ppm relative to Nagaoka, but this degree of accuracy is essentially meaningless. Thus the formula given above is the most suitable for general use on coils that are not too short.

A good formula gives an overview of the shape of the response and allows you to get a *workable* solution. Errors caused by the proximity of metal objects, for example, will swamp the errors in these formulae. Wheeler's formula is useful because it shows that for long coils, the inductance is proportional to the diameter squared. My formula below is accurate to ±0.13%, and it covers the entire range of aspect ratios. (dia = diameter; len = length)

$$\boxed{F = 0.6253 \times \ln\left(1 + \frac{dia}{len}\right) + \left[\frac{0.2110}{0.590 + \frac{len}{dia}}\right] + 0.1370 \times \arctan\left(1.10 \times \frac{dia}{len}\right) \quad \text{nH/mm}}$$

NOTE: To calculate arctan( ), set your calculator to radians, not degrees.

$\frac{\text{length}}{\text{diameter}}$	H (pF/mm)
50	0.58
25	0.29
5	0.081
2.5	0.056
2.0	0.050
1.0	0.046
0.5	0.050
0.3	0.060
0.1	0.096

It might be supposed that the wire needs to be large in order to have a large surface area, and therefore a low RF resistance (the *skin effect*). However, the increased wire size needs to be balanced against the need to keep the turns separated.

Experimental evidence suggests that the self-capacitance of a single-layer solenoid is *not* related to the separation between the turns. The self-capacitance is instead given as: $C = H \times d$ pF, where the diameter, *d*, is given in mm and *H* is a function only of the length/diameter ratio.[9]

The effect of keeping the turns separated is to increase the Q because the *proximity effect* would otherwise increase the losses in the wire. 'Like currents attract' so the current bunches up, increasing the effective resistance of the path.

For a given length/diameter ratio, both the inductance and the self-capacitance are proportional to the diameter. This means the self-resonant frequency is inversely proportional to the coil diameter; a very high frequency inductor therefore needs to be small. The SRF is also inversely proportional to the number of turns, although the turns must still fill the winding space for the formulae to be applicable. For use at microwave frequencies, conical inductors give the best broadband performance. The idea is to put the narrow end of the cone at the more sensitive point in the circuit in order to minimise the stray capacitance to that point. A single Piconics broadband conical inductor, for example, is rated from 40 MHz to 40 GHz.

At frequencies where the wire resistance is limited by the skin effect, the Q of a

[9] R.G. Medhurst, 'HF Resistance and Self-Capacitance of Single-Layer Solenoids', in *Wireless Engineer*, 24 (1947), pp. 35-43, & pp. 80-92.

single layer coil *can* be increased by making the diameter larger, for a given inductance value. It is difficult to say with confidence that the increase is proportional to the square root of the increase in diameter,[10] however, because one has to be careful to use the optimum wire size and spacing.

At lower frequencies (< 200 kHz) the idea of using the minimum amount of wire for a given inductance is both useful and experimentally verified.[11] This suggests a diameter-to-length ratio of 2.5; a short coil. Also ***Litz wire*** is strongly indicated as a method of getting better Q. If you are way off with the diameter-to-length ratio, or the wire spacing is wrong, you could be losing Q by a factor of 2 or more.

Try different coil sizes and aspect ratios to suit your frequency and application. Recommendations for optimum wire and coil size vary with operational frequency so the optimum gap between turns is somewhere between one quarter the wire diameter and two hundred times the wire diameter!

Single-turn loops can be formed inadvertently and it is therefore convenient to be able to estimate the self-inductance created.[12]

$$\boxed{L = 2\times\left(\Delta-\theta+\ln\left[\frac{\textit{loop perimeter}}{\textit{wire diameter}}\right]\right)\ \text{nH/cm of perimeter}}$$

where θ is a shape constant, having values of:

1.065	for a circle
1.250	for a hexagon
1.468	for a square
1.811	for an equilateral triangle
$1+\dfrac{0.5}{0.074+a^{0.6}}-\dfrac{\sin(a\pi)}{10}$	for a rectangle with $\dfrac{\text{long side}}{\text{short side}}=a$; $1>a>0.033$ (rectangular coil approximation accurate to ±0.5%)

Δ is a skin depth correction factor, having a maximum value of 0.25 at DC and falling to zero when the skin depth is a small fraction of the wire radius. This factor accounts for the reduction of internal inductance due to flux within the wire.

When the loop perimeter is very much larger than the wire diameter, the shape of the coil is unimportant. The inductance is primarily a function of the perimeter length, being in the region of 10 nH/cm to 30 nH/cm. The techniques used to derive these formulae are very old.[13]

For a circular loop of 1 mm wire, with a perimeter of 5 cm:

$$L = 2\times5\times(-1.065+\ln[50]) = 28.5\ \text{nH}$$

[10] F.E. Terman, 'Losses in Air-Cored Coils at Radio Frequencies', in *Radio Engineer's Handbook*, 1st edn (New York: McGraw-Hill, 1943; repr. London, 1950), pp. 74-77.

[11] P. Dodd, 'Low Frequency Coil Q by Bill Bowers', in *The LF Experimenter's Handbook*, 2nd edn (Radio Society of Great Britain, 1998), pp. 1.21-1.23.

[12] V.J. Bashenoff, 'Abbreviated Method for Calculating the Inductance of Irregular Plane Polygons of Round Wire', in *Proceedings of the Institute of Radio Engineers*, 15, no. XII (Dec 1927), pp. 1013-1039.

[13] R.G. Allen, 'The Establishment of Formulae for the Self-Inductance of Single-Turn Circuits of Various Shapes.', in *Experimental Wireless and The Wireless Engineer*, V, no. 56 (May 1928), pp. 259-263.

Making the loop 10× bigger:

$$L = 2 \times 50 \times \left(-1.065 + \ln[500]\right) = 515\,\text{nH}$$

Filling a room, a 4 m square loop gives:

$$L = 2 \times 1600 \times \left(-1.467 + \ln[16{,}000]\right) = 26.3\,\mu\text{H}$$

In the formula, the natural log term has to be bigger than the shape constant term in order for the inductance to be positive. The formula is an approximation based on a shape where the wire is thin in relation to the overall shape.

Single-turn inductors are used by physicists to create mega-gauss (>100 T) fields by discharging a capacitor bank into the inductor.[14] Large coils are also used in Audio Frequency Induction Loop (AFIL) systems in Churches, theatres and lecture halls; the sound is used to modulate the current in the loop, enabling hearing aid wearers to use the telecoil pickups [15] in their hearing aids to get clearer sound. Note that this is a base-band modulation system in that a 1 kHz tone is converted to a 1 kHz magnetic field modulation.

One problem with a simple air-cored inductor is that it acts as both a transmitting and a receiving antenna. A metal screening can is therefore useful for preventing unintended transmission or reception; in other words to prevent unintentional interaction with other circuitry. The two key things to decide about metal screening cans are the material and the size to be used. Whilst a "thick" can, by which I mean a can with a wall thickness of at least three skin depths, will give adequate shielding, the material used will affect the resulting Q.

Qualitatively, the shield will always reduce the Q, but the lower the conductivity of the shield, the worse the Q will be. It is impossible to give an analytic expression for the Q reduction, so one can either use the best materials, or experiment to see how much worse the lower conductivity materials are. Thus inductor screening cans are best made from copper or aluminium; brass, steel, and tin-plate being less ideal.

In addition to lowering the Q, the screening-can will also lower the inductance. Qualitatively, the closer the screen is to the coil, the greater the loss in inductance. A simple empirical formula for the loss in inductance has been found for a cylindrical coil of radius r and height h, fully encased by a coaxial cylindrical screen of radius R, when there is a uniform gap, $g = R - r$, between the screen and the coil:[16]

$$\boxed{\frac{\Delta L}{L} = \left(\frac{r}{R}\right)^2 \times \frac{1}{1 + 1.55 \times \frac{g}{h}}}$$

This formula is claimed to be accurate to within a few percent. Using the formula for incomplete screens, or nearby pieces of metal, one could measure the gap and pick a

[14] F Herlach, 'Research with Pulsed Magnetic Fields from the Early Beginnings to Future Prospects', in *Megagauss Fields and Pulsed Power Systems*, ed. by Titov & Shvetsov (Nova Science publishers, 1990), pp. 15-20.

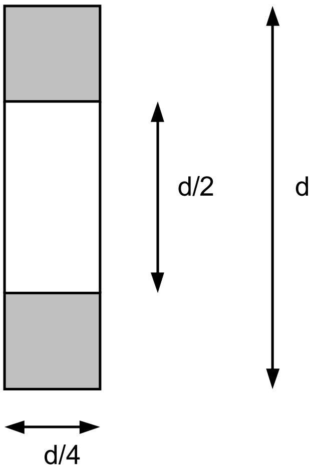
[15] BS EN 60118-4:1998. Hearing Aids- Part 4: Magnetic field strength in audio-frequency induction loops for hearing aid purposes.

[16] A.G. Bogle, 'The Effective Inductance and Resistance of Screened Coils', in *Journal of the IEE*, 87 (1940), pp. 299-316.

value of R such that a circular screen would not be further away than the pieces of metal, thereby giving an upper limit to the inductance loss.

For large air-wound inductors in a small space it is necessary to use a multilayer winding. The maximum inductance for a given amount of wire is given by the *Brooks coil*, published in 1931, a short circular coil, the side cross-section of which is shown below.

FIGURE 7.5A:



Scaling everything to the outer diameter, d, the inner diameter is half this, and the winding area has a square shape of side $d/4$. The resulting optimal inductor has an inductance given by:

$$\boxed{L = 6.37 \times d \times n^2 \text{ nH}}$$

where the diameter d is given in cm, and the number of turns is n.

If heavily insulated wire is used, the inductance is increased slightly at the expense of extra series resistance. Using w as the ratio of the insulated wire diameter to the inner core diameter:

$$L = 6.37 \times d \times n \times (n + 0.115 + 0.739 \times \ln[w]) \text{ nH}$$

The resistance of the coil is obtained from the length of wire used, this being $n\pi$ times the mean diameter per turn. $\boxed{length = 0.75n\pi d}$

For a given number of turns, doubling the coil's diameter doubles the inductance. It also doubles the length of wire used, but it increases the cross-section by ×4, thereby halving the overall resistance. Thus for a given number of turns, the Q increases as the square of the coil's diameter.

One design procedure is therefore to select an overall coil diameter, calculate the number of turns required, then calculate the maximum overall wire diameter. A first estimate would give a wire diameter of $\frac{d}{4\sqrt{n}}$. However, on closer inspection it is seen that layers of circular wire can theoretically be placed at 0.87 diameters apart. This suggests a wire diameter perhaps 5% larger than the initial estimate.

Note that the Brooks coil has considerable self-capacitance, due to its multi-layer construction, and therefore has a much lower self-resonant frequency than a single-layer solenoidal coil of equal inductance.

7.6 Mutual Inductance

When current flows in an inductor, a magnetic field is generated. This flux *links* with the turns in the inductor producing a voltage when the current is changing. The relevant equation is *Faraday's Law* of induction.

$$\boxed{E = L \cdot \frac{di}{dt}}$$

For air-cored inductors, if the turns are pushed closer together the inductance increases, as noted earlier in this chapter. This increase is explained on the basis that more of the flux from one end of the coil reaches the other end; there are more *flux linkages* and hence more inductance.

Any nearby coil may 'intercept' some of the flux, by which is meant that some of the flux from the first inductor may pass through some of the turns in the nearby coil. If a current *i* in coil 1 generates a voltage *V* in coil 2, then the same current *i* in coil 2 produces the same voltage *V* in coil 1. The effect of this is that the circuit model for each of these coils is now an inductance, representing the self-inductance of the individual coil, in series with a *mutual inductance* driven by a current from the other coil.

$V_1 = L_1 \frac{di_1}{dt} + M \frac{di_2}{dt}$ A similar equation applies to the second inductor.

The mutual coupling is positive if the flux from coil 2 increases the flux in coil 1. This gives the standard way of measuring the mutual inductance between two coils. Measure the inductance of both coils wired in series-aiding then swap the connections to one of the coils to measure the inductance of both coils wired in series-opposition.

$$L_{\text{AIDING}} = L_1 + L_2 + 2M \qquad\qquad L_{\text{OPPOSING}} = L_1 + L_2 - 2M$$

There is no confusion as to which measurement is the series-aiding one because it is always the larger of the two. The mutual inductance value is therefore

$$\boxed{M = \frac{L_{\text{AIDING}} - L_{\text{OPPOSING}}}{4}}$$

In a solenoid, the flux from one turn enhances {adds to} the flux in the adjacent turn; the current elements are always going in the same direction. However, in a single-turn loop, the current on the opposite side of the loop is flowing in the opposite direction. Thus the mutual inductance between the parts of the coil reduces the overall inductance. Thus minimising the gap between the *go* and *return* conductors reduces the inductance from the "isolated wire" value.

The mutual inductance between two thin parallel wires is dependant on the ratio of the length to axial separation ratio, $x \equiv \frac{length}{axial\ separation}$, of the wires.

$$\boxed{M = 2 \times \left[\frac{1}{x} - \sqrt{1 + \frac{1}{x^2}} + \ln\left(x + \sqrt{1 + x^2}\right) \right] \text{ nH/cm}}$$

For coaxial cables the mutual inductance between the inner and outer conductors is equal to the self-inductance of the outer conductor. For small-signal applications, the point at which the reactance of the shield (outer conductor) is equal to the resistance of the shield is known as the *shield cutoff frequency*. Above this frequency, any signal voltage developed across the shield will be inductively coupled to the inner conductor. It is this effect which makes scope probes relatively immune to common-mode noise above a few kilohertz provided the ground lead on the probe is of almost zero length.

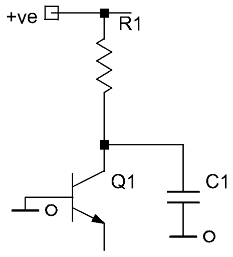
It is found in practice that connecting a grounded signal source to a grounded measuring instrument, such as an oscilloscope, can cause significant low frequency (<1 kHz) measurement noise. This noise has experimentally been found to spread over a 10:1 amplitude range simply due to the exact type of 1 m coaxial cable used for the signal connection.

For aluminium wire armoured (AWA) power cables, the situation is reversed. Power current flows down the inner conductor. If the individual phases of a three-phase power system are running in widely spaced individually armoured cables, there can be circulating currents induced into the armour. The phase current couples to the armour by mutual induction. If the armour is bonded to ground at both ends, there is a circulating current. It is a bad idea to space the phase cables far apart because in this case the circulating current in the armour can get nearly as large as the current in the phase conductors.

7.7 Peaking

The subject of *peaking* goes back to the earliest beginnings of valve amplifiers but is inadequately covered in modern texts. Here is a common sub-circuit:

FIGURE 7.7A:

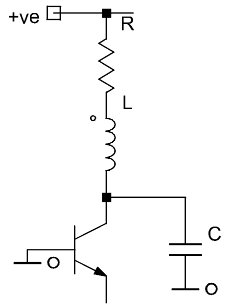


Suppose the signal onto the RC network comes from a common-base stage. This could be a driver for the electrostatic deflection plates of a CRT [Cathode Ray Tube], it could be the video drive for a CRT, it could be a drive for a CCD [Charge Coupled Device], it could even be a level translation scheme between logic families. All the load, ***parasitic*** and ***stray*** capacitance is lumped into C1 and is being driven from Q1. The problem is how to optimise the bandwidth.

More bandwidth is required, without making R1 smaller [as that would dissipate more power]. The lower resistance would also require more gain from the amplifier circuitry, and possibly a larger transistor for Q1. The larger transistor would have more capacitance.

Put an inductor in series with the resistor. It resonates with the capacitor; it stops current being wasted in the resistor on the transient edge; it makes the impedance of the series arm higher so that more current goes into the capacitor.

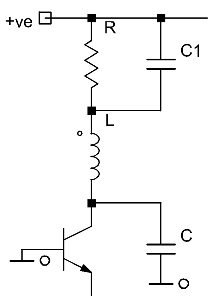
FIGURE 7.7B:



Using a circuit simulator it can readily be established that the value of L for optimum ***monotonic*** frequency response is: $\boxed{L = 0.414 \times CR^2}$

Using this value the bandwidth is increased by a factor of ×1.72. Just adding an inductor has given 72% more bandwidth! The downside is that it gives a 3.1% overshoot on pulses. If a flat pulse response is needed then use $\boxed{L = 0.255 \times CR^2}$. The bandwidth improvement is then 43%. The risetime is improved by 30%.

FIGURE 7.7C:



The addition of one inductor has given 30% faster risetime and 43% more bandwidth; good value for money. The addition of capacitor C1 takes the risetime improvement up to 41%, with 66% bandwidth improvement, and still with no overshoot. The load is still represented by C.

In this case $\boxed{L = 0.5 \times CR^2}$ and $\boxed{C_1 = 0.165 \times C}$

These results assume that there is no bandwidth loss in the device being driven and that it requires a flat frequency or time response. For example, 10% overshoot on a logic pulse would be quite acceptable.
Practically you might tune this circuit up by trial and error because it could be difficult to measure the capacitance directly. Estimate the value of C and start with $L = 0.3 \times CR^2$. You should expect a bandwidth improvement of up to 40%.

For flat pulse response tuning, the risetime improvement is similar to the bandwidth improvement. The question that then comes to mind is just how much improvement could be achieved if the peaking network were made arbitrarily complex. [Notice that the term *peaking* does **not** necessarily mean that the final output peaks in the frequency ***domain*** or overshoots in the time domain.]

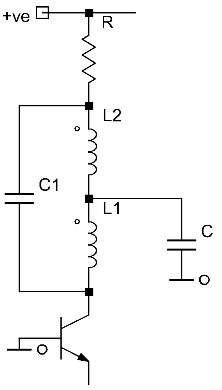
You can always get more bandwidth (frequency domain)
if you accept more overshoot (time domain).

EX 7.7.1: [special interest only] What is the optimum risetime improvement that can be obtained with fixed values of R and C and an arbitrary unspecified linear or non-linear network? (answer on page 110).

Ideally it takes no power to drive a capacitor … if you drive it sinusoidally using an inductor.

In the real world, getting improvement is very much more difficult in the time domain than in the frequency domain. This follows the more complex model, where the frequency domain response can be infinitely improved with a suitably complex network. The reality is that a response is easily tuned in the frequency domain, but the effect of this in the time domain is almost always to give overshoot and possibly also ringing on the pulse response.

FIGURE 7.7D:



If you have a demanding application and the peaking schemes given previously do not give the desired performance then you should consider using a *bridged T-coil* solution.

This looks very good on simulation. The values are:

$$L_1 = L_2 = 0.327 \times CR^2$$

$$k = 0.5\,(\text{couplingcoefficient between } L_1 \text{ and } L_2)$$

$$C_1 = 0.081 \times C$$

This gives a bandwidth improvement of ×2.67

You have 100 MHz bandwidth with R and C, then you peak it to 267 MHz. But this is done without introducing *any* overshoot. The risetime is speeded up by the same factor.

Once you have a solution worked out then it is not going to be expensive to implement. It is primarily the design cost that you need to worry about. The other peaking schemes can be dropped into an existing circuit very easily. This one is going to take considerably more effort because a centre tapped coil with a coupling coefficient of 0.5 is not a standard stock item; you will have to have one made to order. An alternative suggestion for changing the coupling coefficient between the coils is to change their separation or orientation. This might be better for prototyping.

The ability to gain an extra 10% to 20% improvement may not seem important, but can be critical when meeting some sort of ‘external’ requirement. Suppose you have to meet a data rate of 1 Gb/s as a marketing specification. A banner spec of 970 Mb/s is not going to be acceptable. Thus getting the last few percent of speed, accuracy, &c, can really be of benefit.

7.8 Simple Filters

When making a filter for an AC application, there are six key factors:

- the mean insertion loss in the passband
- the insertion loss flatness in the passband
- the bandwidth
- the steepness of the attenuation roll-off
- the ultimate achievable attenuation
- the input and output matching (for 50 Ω or higher impedance systems)

For a time domain application, the pulse response of the filter becomes more important than the AC measurements given above.

An RC filter can never give as good a performance as an optimum LCR filter. The LCR filter will always give a sharper corner and less deviation from nominal in the passband. The trouble is that at frequencies in the kilohertz region, inductors in filters are both expensive and bulky. In this case it is better to make *active filters* using capacitors and opamps. Active filters date back to at least 1955 and the ***Sallen-Key*** filter design.[17]

The steepness of the attenuation roll-off well outside the pass band is governed by the *order* of the filter. A single-RC filter is a first order filter. It rolls off at 20 dB/decade, also expressed as 6 dB/***octave***.

$$\boxed{T = \frac{1}{1 + j\dfrac{f}{B}}}$$

This is the normalised *transfer function* of a *single-pole* system with a 3 dB bandwidth B, where $B = \dfrac{1}{2\pi CR}$ for an RC filter.

The 20 dB/decade roll-off occurs because when $f \gg B$, the imaginary term becomes dominant and the transfer function simplifies to: $\left| T \right| \approx \dfrac{B}{f}$.

7.9 Magnetic Fields

A long straight circular wire carrying a steady direct current produces a simple radially symmetric field pattern. *Ampère's Circuital Law* gives the field.

The magnetic field from an isolated long conductor is inversely proportional to the

[17] R.P. Sallen, and E.L. Key, 'A Practical Method of Designing RC Active Filters', in *IRE Transactions on Circuit Theory*, CT-2 (March 1955), pp. 74-85.

distance from the centre of the conductor, assuming the *return* conductor is infinitely far away. To calculate the actual field when the return conductor is not infinitely far away, use the ***superposition theorem*** and add the fields from the two conductors taken one at a time. It is important to note that this is a vector sum; the directions of the fields must be taken into account.

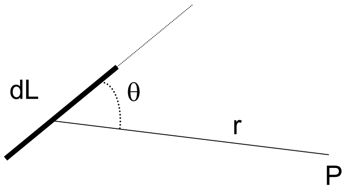
@EX 7.9.1: Two thin parallel conductors supply a resistive load a long distance away. If the current is I and the conductors are a distance d apart, calculate the magnetic field intensity H at a distance r from one of the wires in the same plane as the pair of wires, assuming the wires are surrounded only by air.

@EX 7.9.2: Given the field intensity H (A/m), how do you calculate the magnetic flux density B (Tesla) when the medium is air?

@EX 7.9.3: For a sinusoidally varying, low frequency, uniform magnetic field intensity H (A/m), frequency f (Hz), what is the voltage induced in a single-turn wire loop having an open area of 1 cm^2.

The magnetic field on the axis of a single-turn circular coil can be calculated easily due to its symmetry. The technique is to integrate the field due to small current elements using the ***Biot-Savart Law***.

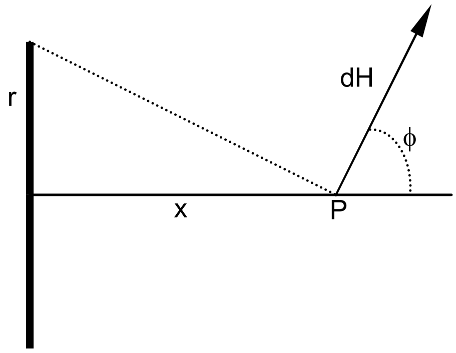
FIGURE 7.9A:



$$\boxed{dH = \frac{I \cdot dL \cdot \sin(\theta)}{4\pi r^2}}$$

The small conducting element, *dL*, carrying current, *I*, produces a contribution, *dH*, to the overall magnetic field intensity, *H*. These elements of field intensity are summed vectorially for the whole current path. The direction of the field intensity is perpendicular to both the direction of the current element and to the direction of the distance vector **r**. In this drawing, *dH* goes into the page (perpendicular) if the current is flowing up the page.

FIGURE 7.9B:



The distance from any small current element in a circular coil to a point P on the coil's axis, a distance x from the plane of the coil, is $\sqrt{r^2 + x^2}$. The current element is always at right angles to the line drawn to the point P, making $\sin(\theta) = 1$ in the Biot-Savart formula.

Using polar coordinates to make the later integration easier, with α as the position of the conducting element within the plane of the circle:

$$dH = \frac{I \cdot r \cdot d\alpha}{4\pi\left(r^2 + x^2\right)}$$

Note that *dH* is the magnitude of the vector field due to the current element, but only the axial component is necessary. All the radial field components cancel during the integration.

The axial component is: $dH\big|_{AXIAL} = dH \cdot \cos(\phi) = dH \cdot \dfrac{r}{\sqrt{r^2 + x^2}}$

Giving $dH\big|_{AXIAL} = \dfrac{I \cdot r^2 \cdot d\alpha}{4\pi \cdot \left(r^2 + x^2\right)^{3/2}}$

Integrate this to get the total field $H = \displaystyle\int_0^{2\pi} \frac{I \cdot r^2}{4\pi\left(r^2 + x^2\right)^{3/2}} \cdot d\alpha = \frac{I \cdot r^2}{2\left(r^2 + x^2\right)^{3/2}}$

$$\boxed{H = \frac{n \cdot I \cdot r^2}{2 \cdot \sqrt{\left(r^2 + x^2\right)^3}}}$$

If there are *n* turns in the coil then, by superposition, the field is *n* times stronger.

According to this formula, when $x >> r$ the magnetic field intensity is $H \approx \dfrac{n \cdot I \cdot r^2}{2 \cdot x^3}$, an inverse cube law with distance. For $x > 6r$ the approximation is accurate to better than 5%. Unwanted magnetic interactions are therefore rapidly reduced by increasing the spacing from the emission source.

Expressions for the magnetic field in the plane of a circular coil are seldom seen due to their mathematical complexity. Integration of the Biot-Savart formula for the field in the plane of a circular coil gives:

$$\boxed{H = \frac{I}{2\pi r} \cdot \int_0^{\pi} \frac{1 - a \cdot \cos(\theta)}{\sqrt{\left(1 + a^2 - 2 \cdot a \cdot \cos(\theta)\right)^3}} \cdot d\theta}$$

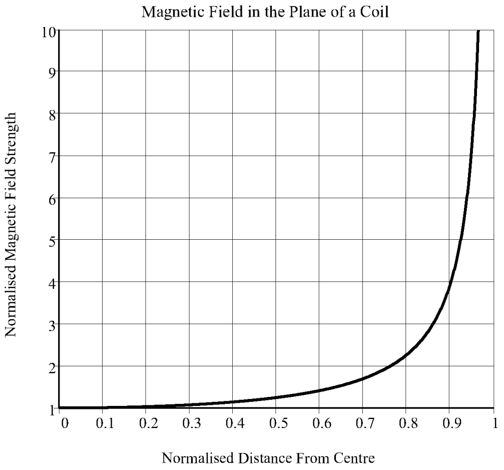
… where *a* is the normalised distance from the coil's axis.

a=0 is on the axis and *a*=1 is at the coil winding itself. Advanced text book solutions to this integral involve answers in either power series form, or in terms of standard complete ***elliptic integrals***. This non-standard elliptic integral is however easily solved by numerical integration on a computer. Unfortunately, none of these methods gives a feel for the result or a simple formula for plotting the curve.

The approximation formula below is accurate to ±0.4% relative to the integral formula (above):

$$\boxed{H \approx \frac{nI}{2r}\left[1 + \frac{a^2}{\pi} \cdot \left(\frac{2 - a}{1 - a}\right) + \frac{0.0092 \cdot a}{1.008 - a}\right]}$$

FIGURE 7.9C:



The field is seen to become infinite when $a = 1$, a result of assuming that the wire has negligible diameter. Even the 'exact' formulation therefore becomes inaccurate as the edge of the coil is approached.

Nevertheless, the important point is made that the field near the edge of the coil will be more than 10× larger than that at the centre of the coil.

Whilst the field in the plane of a thin circular coil is highly non-uniform, the field in the cross-section of a long cylindrical coil {a ***solenoid***} is constant. The **H**-field in a long cylindrical coil is always parallel with the axis and is constant across the cross-section. A useful application of this idea is that when a long copper pipe is being used as a magnetic shield (>50 kHz), the axial field is relatively constant within the cross-section. Regardless of the cross-sectional shape of the (long) coil, the field is still constant across the whole cross-section. For a long solenoid, $\frac{L}{r} > 10$, the field intensity in the middle is

$\boxed{H \approx \frac{nI}{L}}$ (with less than 2% error)

Another method of producing a fairly uniform field is by the use of *Helmholtz coils*. In this case a pair of identical thin coils are placed on the same axis with their planes parallel and separated by a distance equal to their radius. The winding direction and the flow of the current are made the same for both coils, allowing the fields to add. The idea is that moving further from one coil gets you closer to the other, the effects cancelling to a limited extent.

The Biot-Savart Law only applies to strictly static magnetic fields. If the current is changing periodically, as $I \cdot \sin(\omega t)$, additional terms are needed in the formula:

$$\boxed{dH = -\frac{I \cdot dL \cdot \sin(\theta)}{4\pi r^2} \cdot \sin\left(\omega\left[t - \frac{r}{c}\right]\right) + \frac{I \cdot dL \cdot \sin(\theta)}{2\lambda r} \cdot \cos\left(\omega\left[t - \frac{r}{c}\right]\right)}$$

The $\left(t - r/c\right)$ factor within the sine and cosine terms takes into account the finite propagation delay of the electromagnetic wave, giving *retarded* values. The first part of the overall formula is the inductive Biot-Savart term, falling as an inverse square law with distance. The second part is the radiation term, falling inversely with distance. Neglecting the phase shift, these two terms are numerically equal when $r = \lambda/2\pi$.

At large distances from any type of electromagnetic source, both the electric and magnetic fields reduce with the reciprocal of distance, a $1/r$ law. This is the *radiation field*, with the ratio between the electric field and the magnetic field defined by $Z_0 = \dfrac{|\mathbf{E}|}{|\mathbf{H}|} = \sqrt{\dfrac{\mu_0}{\varepsilon_0}} \approx 377\,\Omega$. Note that $Z_0 = 120\pi$ for the characteristic impedance of free space is not an exact expression, but is accurate to 0.07%; more than adequate for general use.

@EX 7.9.4: A long distance from a dipole antenna, the electric field intensity is 7.3 mV/m. What is the strength of the magnetic field intensity at this same distance? (The answer is on page 110, at the end of this chapter).

EX 7.9.5: A small unloaded rectangular loop antenna is used to measure the field of an incoming electromagnetic plane wave from a distance source. The loop has a height *a* in the direction of the electric field and a width *b* in the direction of propagation of the wave.

a) Is the loop optimally oriented to receive the electric field?
b) What is the amplitude of the received signal calculated from the electric field? (Hint: You must use retarded values.)
c) Is the loop optimally oriented to receive the magnetic field?
d) What is the amplitude of the received signal calculated from the magnetic field?
e) What is the actual received signal amplitude?

It is often stated that toroidal inductors and transformers produce minimal external magnetic fields. This is not true even when high permeability cores are used. The conventional winding produced by a toroidal coil winding machine “progresses” as it is wound around the core. On average the effect is to make a stray equivalent circuit of a single-turn passing through the middle of the winding, in the plane of the toroid. A toroidal inductor is therefore susceptible to stray flux passing through the plane of the toroid and, reciprocally, it generates stray flux perpendicular to its own plane.

Since the winding gives the equivalent of a single-turn loop, for minimal stray magnetic field the current should enter and leave the winding near the same point.

Whilst some reduction of the stray magnetic field can be achieved by running one of the output wires back around the edge of the toroid, forming a cancelling single-turn loop, better winding techniques exist.[18] One method is to wind alternate layers with the winding progressing in opposite directions. This is easiest to describe on a solenoid. First wind from left to right, then wind from right to left. The progression of the winding head of the winding machine reverses, but all the turns are still clockwise.

[18] B.P. Kibble, and G.H. Rayner, 'Ch 4: Transformers', in *Coaxial AC Bridges* (Bristol, UK: Adam Hilger, 1984).

7.10 Transformers and Inductive Voltage Dividers

Four specific types of transformers are to be briefly discussed:

1) ***Mains*** transformers (50 Hz/60 Hz)
2) Switched-mode Transformers
3) RF transformers
4) Metrology transformers and ratio devices

Mains transformer design considerations include input-to-output isolation, size, cost, efficiency, and no-load to full-load regulation {volt drop on load}. Optimisation of any one property will tend to worsen several of the other properties. Making a transformer larger makes it more efficient, but also more expensive and heavier. Using a split-bobbin construction, with primary and secondary windings axially separated along the core, makes the input-to-output isolation better, but worsens the no-load to full-load regulation {volt drop}.

Mains transformer cores are made from grain-oriented silicon iron, laminated parallel to the magnetic flux lines to minimise eddy current losses. Such a transformer takes a no-load current, the *magnetising current*, which is distinctly non-sinusoidal when connected to a sinusoidal supply. This non-sinusoidal current is a consequence of the B-H curve of the silicon iron.

An unloaded transformer is an inductor. When this inductor is connected across a supply rail the inductor therefore draws a current due to its finite reactance. However, mains transformer designs are rated for the full-load current rather than the magnetising current. A given size of core will support a certain maximum power output, a bigger core being needed for more power.

By making simplifying assumptions of the no-load condition of the transformer, a very important equation can be developed. Using Faraday's Law of Induction, the applied primary voltage is related to the rate of change of flux in the core, ultimately giving the peak flux density in the core.

$$V = N \cdot \frac{d\phi}{dt} = N \cdot \frac{d(B \cdot A)}{dt}$$ where V is the primary voltage, N is the number of turns on the primary, ϕ is the flux in the core, B is the flux density in the core and A is the cross-sectional area of the core. Because the primary voltage is sinusoidal, the flux is required to be sinusoidal. Converting from RMS to peak flux gives:

$$V = \frac{\hat{B}}{\sqrt{2}} \cdot A \cdot N \cdot \omega = \frac{2\pi}{\sqrt{2}} \hat{B} \cdot A \cdot N \cdot f \qquad \text{then} \qquad \boxed{V = 4.44 \cdot \hat{B} A N f}$$

… "the transformer equation". The number of turns on the primary is governed by the maximum allowable peak flux density in the core. Silicon iron can tolerate up to 2 T before saturation, but even on peak mains it is advisable to reduce this to below 1.5 T in order to minimise the core loss.

An unloaded transformer is a (saturable) inductor and the current drawn is therefore ideally 90 degrees out of phase with the supply voltage. If the supply is connected at peak mains there is no AC transient as the current starts from zero, just like the steady state. It turns out that if the supply is connected at the zero crossing point of the supply this gives the worst possible AC transient. The core flux will try to reach double its steady-state maximum value, almost certainly saturating the core and giving a large surge current. Any fuse in the circuit therefore needs to be rated to withstand this switch-on surge.

Note that the power handling capability of any particular core is not given by an equation. A certain number of turns are required on the primary to keep the magnetising current down. This, and the desired secondary voltage(s) then set the number of turns on the secondary winding(s). The heaviest gauge copper wire is used that just fills up the winding space and the design is finished.

The "power handling" capability of the core is seen to be empirically derived from the power handling capability of transformers wound on the same core. These figures are then given in the manufacturer's data sheets or application notes. Ordinarily the electronics designer would design the transformer only in terms of input, output and efficiency requirements. The rest would usually be done by the manufacturer's in-house specialist.

The biggest problem with mains transformer based power supplies for equipment is that the input voltage range for world-wide operation is so high. On the 230 V range, modern equipment is expected to run from 190 V to 260 V without having to change (winding) taps on the transformer. It is for this reason that switched-mode power supplies are so popular. They can handle this input range without becoming inefficient. If a switched-mode supply produces too much magnetic or switching noise for your application, a ferro-resonant transformer (constant voltage transformer) may be the answer, albeit at increased cost. These transformers can reduce an input voltage range of ±15% to an output voltage change of ±3% without the excessive power loss of a linear voltage regulator. The result is greater efficiency than a standard transformer in the same application, greater reliability than a switched-mode supply, and also less electrical noise than a switched-mode supply.

Because switched-mode transformers are run at 1000× the frequency of mains transformers, the core material needs to be ferrite in order to minimise the core losses. Note that ferrites saturate at something like 200 mT, a factor of 10× lower than silicon iron. Nevertheless the size of a transformer scales roughly in proportion to the operating frequency, making switched-mode supplies both smaller and lighter than equivalent mains power transformers.

Whilst ***skin effect*** and *proximity effect* are less relevant at mains frequencies, they have increasing significance above 10 kHz. Windings on switched-mode transformers may only consist of two or three turns, but these turns have to be wound as multiple strands in parallel, or as copper strip, in order to keep the copper losses down.

Transformers for RF purposes may consist of open coils on plastic formers, a *tuning slug* being used to adjust the mutual coupling. Typically such a transformer would have an overall metal screen to minimise stray capacitive couplings to the windings. Nevertheless capacitive coupling within the transformer would be significant. Whilst a circuit diagram may show a single-ended signal applied to the transformer primary and a balanced secondary winding, the parasitic capacitance of the windings could easily make the output unbalanced. Thus inter-winding electrostatic screening may be advisable. Alternatively, the transition from unbalanced to balanced signals may need two transformers cascaded to improve the final balance, the amount of imbalance being reduced by each of the transformers.

Transformers used for metrology applications are not required to supply significant amounts of power. Nevertheless the core losses must be minimised in order to make the input impedance acceptably high. For this reason, and to improve the magnetic coupling between windings, these precision transformers often use toroidal cores wound of ultra-high permeability material (supermalloy) in the form of a tightly wound tape, geometrically much like a reel of sticky tape.

Since the number of turns in an inductor is an unchanging quantity, inductive voltage dividers can be more stable than resistive dividers, providing the parasitic errors are minimised. “8-dial” inductive voltage dividers (IVD) are commercially available that have 0.01 ppm resolution and better than 0.1 ppm accuracy within the frequency range of 100 Hz to 1 kHz.[†]

****

ANS 7.7.1:

The best risetime would be achieved if the resistor were to be removed until the capacitor had charged to its full extent. If the input current step is I and the resistor is R then the aiming voltage for the pulse is I·R. The voltage on the capacitor must slew from 0 to I·R, the risetime being 80% of this time when using the standard 10%- 90% risetime rule. The full slew time is found from $I = C \cdot \frac{dV}{dt}$

giving $\Delta t = C \cdot \frac{\Delta V}{I} = C \cdot \frac{I \cdot R}{I} = CR$

The risetime using this switched resistor scheme is therefore 0.8×CR

The risetime of the unmodified RC network is 2.197×CR

The best risetime improvement possible is therefore 2.75×

The inductor solution does not look so wonderful now that you see what is possible. There are networks that get much closer to this optimum, but they involve centre tapped inductors, T-coils.

ANS 7.9.1: see text

ANS 7.9.2: see text

ANS 7.9.3: see text

ANS 7.9.4:

This is stated as a ‘far field’ problem, albeit in a round about fashion. The electric and magnetic fields are related by the wave impedance, this being the characteristic impedance of free space in the far field.

$$|H| = \frac{|E|}{Z_0} = \frac{7.3\ \text{mV/m}}{377\Omega} = 19.4\ \mu\text{A/m}$$

[†] Eg. Tinsley 5560K 8 dial inductive voltage divider.

CH8: the diode

8.1 Historical Overview

There are lots of different types of diodes, with a very broad range of prices. An ordinary low-current silicon diode costs around $0.02. Step-recovery diodes and tunnel diodes can be upwards of $50; mm-wave mixer diodes can cost $1000 each.

'Diode' is formed from the Greek root *di-* meaning two, and the ending of electr*ode*. The first true diodes were *thermionic diodes* {valve diodes; vacuum tubes} patented by Fleming in 1904. The *cat's whisker diode*, a piece of wire pressed into a natural crystal of lead sulphide or silicon, was patented by Pickard [1] in 1906 as the *crystal detector*, but other radio frequency detectors had been around since at least as early as 1899.

Earlier rectification methods were not deeply explored. Braun, for example, reported up to 30% deviation from Ohm's law when changing amplitude and direction of the current through various crystals in 1874. Electrolytic rectification was discovered at least as early as 1857, its use for the rectification of alternating currents being reported in 1891.

The copper oxide rectifier was first developed in 1922, but not widely reported until 1927. By 1935 operation up to 100 kV was quoted for electrostatic dust precipitation equipment, along with operation at 12,000 A for electroplating equipment.[2] Copper oxide rectifiers were also used in moving coil multi-meters for many years (up until the late 1980s) to rectify the current for the AC ranges.

When copper gets oxidised, the resulting contact can rectify to a greater or lesser degree. This effect needs to be considered in low noise and/or low distortion circuitry. Even "power" applications are not immune to the effect. Carbon brushes bearing on copper commutators produce excessive EMI as a result of the oxide layer on the copper. Chromium plating the commutator reduces this EMI by a factor of ten!

The selenium rectifier was introduced in Europe in 1928 and by 1941 it was estimated that 88% of metallic-junction rectifiers (outside of the USA) were of the selenium type. Selenium displaced copper oxide because the forward resistance of the selenium parts was lower and therefore their efficiency was higher. The selenium rectifier had a forward resistance of around 9 Ω/cm² of junction area. However, it could only withstand 25 V peak reverse voltage per junction.[3] This made the rectifiers large when used for high currents and/or high voltages. For high power use, the steel plates which formed half of the junction acted as the fins of a naturally convected heatsink.

Another difficulty with both copper oxide and selenium rectifiers was the ageing characteristic. The forward resistance of the rectifier was found to increase by up to 50% after 10,000 hours of operation. Thus when the germanium rectifier appeared in the early 1950s, with its much lower forward resistance and a negligible ageing characteristic, it was quickly adopted. However, silicon rectifiers came on the market only a few years later and their superior reverse leakage characteristics meant that germanium rectifiers

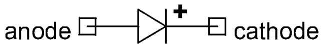
[1] 'Greenleaf Whittier Pickard', CDROM edn (Encyclopaedia Britannica, 2000).

[2] A. Arnold, 'Copper Oxide Rectifier', in *The Modern Electrical Engineer*, vol II of IV, 2nd edn (London: Caxton Publishing Company, 1935), pp. 213-214.

[3] C.A. Clarke, 'Selenium Rectifier Characteristics, Applications and Design Factors', in *Electrical Communication*, 20, no. 1 (1941), pp. 47-66.

were also quickly superseded.

FIGURE 8.1A:



Here is the standard diode symbol with an additional *plus* sign. The + sign is not part of the symbol, but is occasionally seen on power supply circuit diagrams; it is intended to indicate where the positive supply comes out.

The terminal names anode and cathode in a diode are a hangover from the days of valves. Originally, however, the names *electrode*, *anode*, *cathode* were proposed by Faraday for use in electrolysis (1834).[4]

On the symbol above, conventional current flows easily from left to right, ie from anode to cathode. This is indicated by the "arrow" formed by the body of the symbol. I shall neglect thermionic diodes completely; any use they may still have must be very specialised.

8.2 The Silicon Diode.

The silicon diode is the main type of diode that would be used in circuit design. It is the most common, the most readily available, and the cheapest.

The 'diode equation' gives the diode current I_D as a function of the applied voltage V_D …

$$\boxed{I_D = I_S \cdot \left(\exp\left[\frac{V_D}{\eta V_T} \right] - 1 \right)}$$

Where:

$V_T = \frac{kT}{e}$, the thermal voltage, ≈25 mV (room temp)

η (eta), the *ideality factor*, is between 1 and 2.

I_S, the *reverse saturation current*, which approximately doubles every 10°C.

(Limits for this figure are not given in manufacturer's data sheets.)

$$\frac{dI_D}{dV_D} = I_S \cdot \exp\left[\frac{V_D}{\eta V_T} \right] \cdot \frac{1}{\eta V_T} \approx \frac{I_D}{\eta V_T}$$

Differentiating the diode equation gives the small-signal equivalent resistance.

$$\boxed{R_D = \frac{dV_D}{dI_D} = \frac{\eta V_T}{I_D}}$$

Unfortunately η changes with diode current. Empirically, η reduces to 1 at high currents.

It is undesirable for the small-signal resistance of a diode to be large and it is therefore desirable that the ideality factor, η, should be 1. When used on RF/mm-wave mixers, a larger η results in worse noise ($noise \propto \eta \times T$) and larger conversion loss.

By calculation, a diode running at 100 mA has a small-signal resistance of ≈0.25 Ω, but this figure neglects the *bulk resistance*, and ohmic contacts to the silicon. These give an extra resistance in series with the calculated small-signal resistance. Unless otherwise specified, it is useful to assume a value between 2 Ω and 10 Ω for this bulk resistance on small-signal diodes. Assume that an otherwise unspecified silicon diode has a small-

[4] M. Faraday, *Experimental Researches in Electricity* (Taylor & Francis, 1839; repr. Dover, 1965), paragraphs 662-663, Vol I.

signal forward resistance approximately given by: $R_D = \left[3 + \frac{26}{I_{D(\text{mA})}}\right]\Omega$, at room temperature.

If a diode costs only $0.02 then the manufacturer is unlikely to characterise it extensively. The data sheet will give some huge leakage current at an equally huge voltage, tempting you to 'curve fit' this one data point to the diode equation and thereby estimate the current flow at lower voltages. You will be misled if you do this; an ordinary diode has a poorly defined shunt leakage resistance in addition to the reverse saturation current defined by the diode equation.

The reverse leakage current can also be strongly affected by light, particularly for translucent glass bodied devices. If light manages to get onto the junction, extra charge carriers will be generated, and the leakage current will be dramatically increased. In fact early photo-transistors were ordinary transistors without the standard black paint on the glass body.

For this reason 'low-leakage' wire-ended diodes have black bodies, although strong sunlight may still have some effect. The only way to be sure is to try blocking out the light and seeing if the circuit is affected. Don't assume that a black plastic packaged device is guaranteed to be light proof; test it. Use a non-conductive object to block the light. Conductive items like hands are not good for this sort of test; they can disturb electric fields and cause a shift in a circuit by other mechanisms than the lighting effect being tested for.

8.3 The Germanium Rectifier.

Germanium has a lower volt-drop than silicon for P-N junctions (base-emitter junctions of transistors and diode forward volt drops). However, the leakage currents found with germanium devices were always much worse than for silicon devices so germanium technology has faded away.

The Germanium Power Devices Corporation made the G30R4 up to the early 2000's. This 30 A germanium diode gave a 0.3 V drop at 10 A and 100°C, a considerable improvement on a silicon diode. However, schottky technology is being improved all the time, so that (silicon) schottky diodes out-perform germanium diodes. For example, an IXYS DSSK70-0015B 70 A diode run at 125°C gives only 0.33 V drop at 35 A. For the best efficiency it is necessary to run the devices hot and to use them at less than half their rated maximum current. These may seem like conflicting requirements: less current but higher temperature. A moment's thought reveals that the key is to use less heat-sinking, a cost saving in itself!

There is talk of making germanium schottky diodes for even lower volt drops, but the primary use for germanium nowadays is as an alloy of silicon and germanium, used to get more f_t in bipolar transistors (f_t > 40 GHz). You will see this alloy written as SiGe. Lower voltage-drop rectifiers for high-efficiency switched-mode supplies are now routinely made using *synchronous rectifiers* (described later in this chapter). These synchronous rectifiers are more efficient than the best germanium or schottky devices.

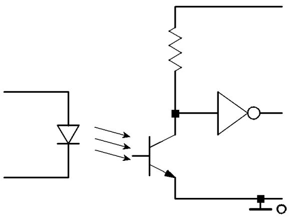
8.4 The Light Emitting Diode (LED)

An LED is the standard way of giving simple front panel indications of the status of a system or a piece of equipment. Lighting LEDs for front panel indications is relatively trivial, but be aware that different colours of LED can have wildly differing on-state voltages. Blue LEDs have typical forward volt drops of 4 V, as compared to ≈1.5 V for red LEDs. The more demanding applications for driving LEDs come when they are being used for isolated interfaces and high-speed data links.

A figure of merit for an analog opto-coupler is the *current transfer ratio*, CTR, the ratio of output current in the photo-transistor to input current in the LED. The LED effectively provides a base current for the photo-transistor. Since the transistor has current gain, the CTR is not restricted to being less than 100%.

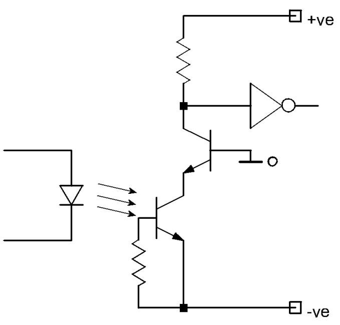
An opto-isolator is a relatively inexpensive way of giving kilovolts of isolation to a digital control line or even an analog signal. Consider a low-speed opto-isolator for a logic line. This might consist of a photo diode at the sending end, optically coupled to a photo-transistor at the receiving end. These could be in one package or separated by some sort of light pipe.

FIGURE 8.4A:



This circuit is slow because the transistor has no base pull-down resistor to remove stored charge when the transistor is required to switch off. The circuit is also slow because of the ***Miller feedback*** of the collector-base capacitance. This digital opto-isolator is a relatively inexpensive interface, but it is not useful above a few kilohertz.

FIGURE 8.4B:

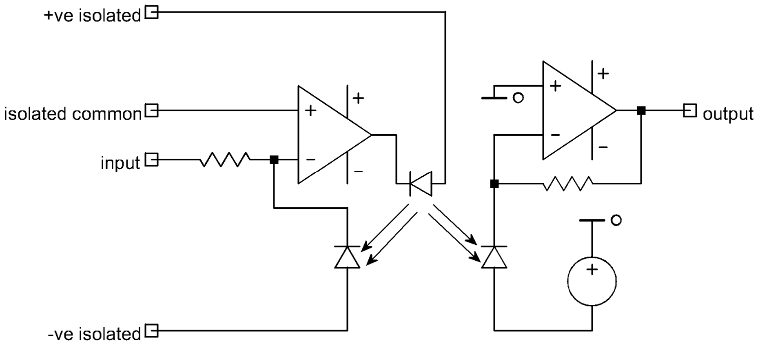


The circuit can be speeded up by the use of a base-emitter load resistor and by feeding the collector into either a ***cascode*** stage or the virtual earth of an opamp

If you are just doing a simple digital interface then you should buy an 'off-the-shelf' digital opto-isolator, rather than make your own. The digital opto will be specified in terms of a minimum input [LED] current and a guaranteed operating temperature range. This considerably reduces the design time, design cost, and risk. However, you may still need to make your own if you need a more sensitive or a faster interface than that provided by off-the-shelf parts.

Direct use of any sort of opto-coupler for transferring an analog signal is out of the question. The non-linearity, variation with temperature, and variation with time, make the system unacceptable, even with specs of a few percent. What is done is to use matched pairs of opto-devices. By far the best is the single LED transmitter driving two photo-diode receivers. If the light is split evenly between the two photo-diodes, then the matching of these photo-diodes becomes the limiting factor.

FIGURE 8.4C:



In a typical configuration, the signal is converted into a defined photo-diode current using feedback around an opamp. A representative device is the Infineon IL300. The key factor for achieving good matching is to operate both photo-diodes with near identical bias conditions.

Operation with reverse bias is *photoconductive* operation. Operation with zero bias voltage is *photovoltaic* operation, which gives better linearity, but lower bandwidth than photoconductive operation. Whilst 12-bit linearity and 30 kHz performance are suggested for photovoltaic operation, the linearity is 10× worse and the bandwidth is 3× better for photoconductive operation.

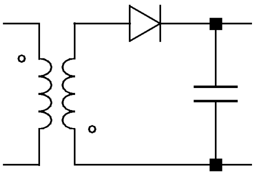
Low-speed (20 kHz) opto-isolators are very inexpensive ($0.50). If you need to go considerably faster and/or with >50 kV isolation then fibre optic isolation may be necessary. 2Gb/s transfer at MV isolation is then achievable, but the cost will have risen from $0.50 to $150.

Transfer of digital signals across isolation barriers is also done by two other methods, one using transformers and the other using pairs of picofarad capacitors driven differentially. All schemes suffer to some extent from fast slewing input signals on the isolated input. The capacitive coupling scheme is, however, particularly sensitive. This effect becomes important for serial data streams referenced to mains signals. The mains voltage has all sorts of large fast transients on it. These give high slew rates which can therefore corrupt the serial data stream.

8.5 The Rectifier diode.

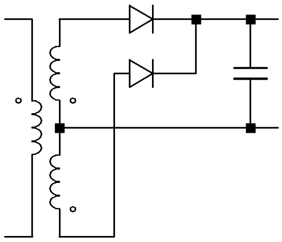
All diodes are rectifiers, but when using the term 'rectifier' the usage in a power application is being emphasised. Small power supplies using a transformer and rectifier(s) are simpler, cheaper, and more reliable than switched-mode supplies.

FIGURE 8.5A:



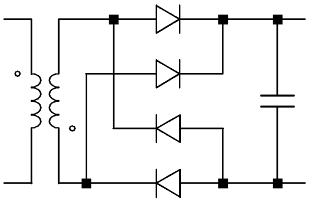
The half-wave rectifier is a very inexpensive way of getting a low current supply (let's say <50 mA). If a lot of current (>100 mA) is required then use a full-wave rectifier; the capacitor can then be halved for a given ripple voltage.

FIGURE 8.5B:



The *bi-phase bridge* gives a full-wave rectified output by using an extra diode and an extra tap on the transformer. Neglecting the transformer efficiency, the bi-phase bridge can more efficient than a full bridge at low output voltages (< 3 V) because there is only one diode in the conduction path at any time.

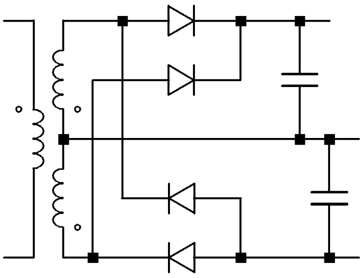
FIGURE 8.5C:



The standard *bridge rectifier* circuit. Do not use this arrangement to produce <2 V DC power rails; the diode drops make both the efficiency low and the voltage stability poor.

A component with 4 diodes, wired up as shown with two input and two output leads, is known as a bridge rectifier.

FIGURE 8.5D:



This circuit could be looked at as a centre-tapped bridge rectifier, or as a pair of bi-phase bridges. It is the standard way of producing dual polarity supplies.

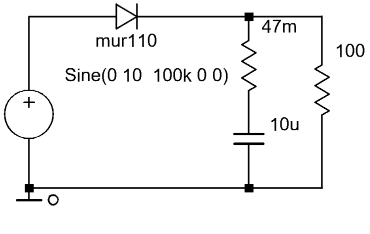
Note that a diode ideally has an almost constant voltage drop across it due to the exponential relationship between volt drop and forward current. Hence the power dissipated is not calculated from the RMS current, but something closer to the mean current. As an example, the STPS20L45C† data sheet gives the power dissipation as:

$$P = 0.28 \times I_{F(AV)} + 0.022 \times I_{F(RMS)}$$

For switched-mode power supplies it is necessary to specify ‘ultra-fast recovery’ rectifiers. The diode is initially conducting during part of the switching cycle. The voltage is forward, the current is forward. When the driving voltage reverses, the current would ideally just stop. Unfortunately real diodes conduct substantial current in the reverse direction. The speed with which this reverse current stops is determined by the *reverse recovery time*.

An ordinary mains frequency rectifier diode is particularly bad at reverse recovery. If placed in a switched-mode power supply, it would overheat, despite being well within its mean and peak current ratings. The efficiency of the power supply would also be noticeably reduced. When the driving voltage reverses, the diode current initially reverses as well, taking energy out of the power rail and dissipating it as heat. Fortunately ultra-fast rectifiers are available with 50 ns reverse recovery times. This means that if they were conducting in the forward direction, and the driving voltage quickly reverses, the reverse current will only flow for around 50 ns.

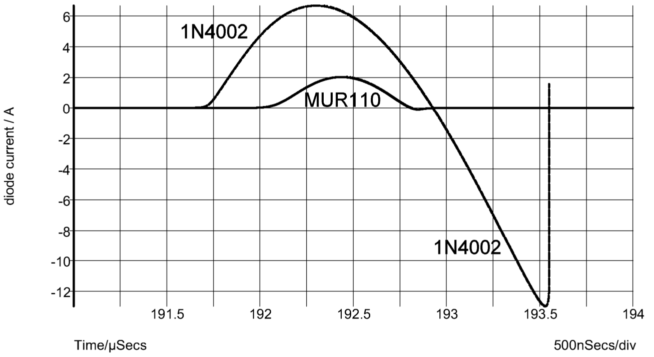
FIGURE 8.5E:



This is a simplified simulation model of a switched-mode power supply running at 100 kHz. The current waveforms in D1 show the whole story.

† STMicroelectronics STPS20L45C data sheet, rev 4, 2007.

FIGURE 8.5F:



The ultra-fast rectifier, MUR110 gives the simple pulsed current waveform. Use of a standard power frequency diode, 1N4002, shows the reverse recovery problem.

The reverse current shown in the simulation is due to the reverse recovery time of the 1N4002 being so slow. The power diode would burn out in this application. The best verification safeguard for this sort of circuit is to stick a thermo-couple onto the body of the diode to make sure it doesn't get too hot. This technique is the best way of ensuring that the diode is not being overrun. Current probes can give false reading in such applications due to the high dV/dt signals also present.

Notice the vertical rise of the reverse current back up to zero on the graph. This *snap recovery* can be very fast and generates EMI as a result. For use on switched-mode power supplies this sort of characteristic would be electrically noisy. The diodes that are optimised to slow down this reverse recovery transient are known as *soft-recovery* diodes. Thus it is highly desirable to use a diode that recovers from the reversed input quickly, but in a gentle manner. This characteristic maximises the power supply efficiency, whilst at the same time minimising the EMI generated.

8.6 The Schottky Diode

Rather than a PN junction, the schottky diode uses a metal-semiconductor junction and operates on majority carrier current flow rather than minority carrier flow. For this reason it is supposed to have minimal forward and reverse recovery times. It has a much lower volt drop than silicon at low current levels, but it is quite possible to end up with more volt drop in a schottky diode than in a cheaper silicon diode.

Schottky diodes get difficult to manufacture at reverse voltages of more than 50 V or so. Thus whilst it is possible to get power schottky rectifiers with a reverse breakdown voltage of 100 V, the forward volt drop can get unreasonably high, >0.85 V [International Rectifier 11DQ10]. To get a lower volt drop it is necessary to use a larger diode.

If you run a 1 A schottky diode at 1 A, the volt drop will be unacceptably high. Manufacturers rate their diodes for the thermal limit rather than the voltage limit. Thus you may have to use a 3 A diode at 1 A to keep the volt drop acceptable. If the volt drop in the schottky exceeds 0.7 V you may be better off using an even larger [higher current rating] silicon diode. This will give the same or better static performance at less cost. In a switching application, it is not just the static loss (forward volt drop) that is important, the reverse recovery characteristic may dominate the power loss.

There is a process-dependant factor for schottky diodes known as *barrier height*, which might typically lie in the range of 0.7eV to 0.9eV. For an ideal schottky barrier, the forward volt drop is proportional to the barrier height. Unfortunately the reverse current increases exponentially as the barrier height is reduced. Thus in a switching situation, the power loss due to reverse current conduction can exceed the forward conduction power loss if a low barrier device is used. This is particularly true in higher voltage applications (say greater than 45 V).

These statements apply to silicon schottky diodes. Up until recently this has been all that has been on offer. As of 1999 you can now get GaAs schottky diodes for >150 V applications. For example an IXYS DGS10-025A can handle 250 V. It has a shorter reverse recovery time and a lower forward volt drop than a conventional ultra-fast silicon rectifier. Basically this is a continuously changing area and you need to look in the catalogues or on the Internet every time you start a new design to see what new and exciting parts are available. Using key components from a 5 year old catalogue is not going to make your designs state-of-the-art.

As of 2001 the latest greatest commercially available technology was silicon carbide (SiC). Silicon carbide has 10× greater breakdown field strength than silicon, can tolerate temperature well in excess of 400°C, and has 3× the thermal conductivity of silicon. The reverse leakage current approximately doubles every 30°C, rather than silicon's doubling every 10°C. The only down-side is that the forward voltage knee is more like 1 V rather than the 0.6 V of silicon. However, this static loss is not a problem for high voltage applications. SiC schottky diodes can achieve 600 V reverse breakdown levels (Infineon SDP04S60, discontinued in 2006 due to lack of demand!) with minimal reverse recovery, and this can greatly enhance the performance of switched-mode power supplies.

SiC is an excellent material for power semiconductor applications, but is quite expensive due to the low volume and increased cost of production. The greater operating temperatures are not being exploited because the packaging technology has not kept step with the semi-conductor technology. Also, standard FR4 PCB material is not suitable for components running at a case temperature of over 200°C, 120°C being a more usual maximum surface temperature for FR4.

Low barrier schottky diodes are used to detect RF signals up into the GHz region. You can simply apply microwave power from a small pickup loop (say 2 cm in diameter) through a low barrier schottky diode into a hand-held DMM to get a measure of the RF field strength. GaAs schottky diodes packaged as beam lead devices can act as simple detectors up to more than 20 GHz, with smaller devices being used out beyond 100 GHz.

When used on power levels below –20 dBm a schottky diode acts as a power detector, which is to say the output voltage is proportional to the input power. At these low input power levels the diode has to drive a high valued load resistance (1 MΩ) in order to give the largest signal.

A schottky diode is a metal to semiconductor junction, but the semiconductor can be doped as either *n*-type or *p*-type. Since *n*-type material uses electrons as the majority charge carrier, and since electrons have a higher mobility than holes, for the highest possible frequency of operation it is preferable to use a metal to *n*-type junction. This allows mixers to be made that operate above 1000 GHz (1 THz).

8.7 The Zener (Voltage Regulator) Diode.

Voltage regulator diodes come in two distinct types as far as physicists are concerned. Designers just group them together as 'zener' diodes. Both types appear as ordinary silicon diodes in the forward direction, and in the reverse direction they both have a well defined, non-destructive breakdown characteristic.

Diodes which breakdown at <6 V use the *zener effect*, which has a negative temperature coefficient. Above 6 V the dominant mechanism is *avalanche breakdown*, which has a positive temperature coefficient.

Because the temperature coefficient changes from positive to negative around 6 V, it is possible to obtain diodes with nominally ‘zero-TC’ in this region, meaning TCs of ±5 ppm/°C [1N829A @ $2.00 each.] The TC is also a function of the operating current. 1N829A diodes are specified at 7.5 mA, but in the range 6.0 mA–9.0 mA a more accurate ‘zero-TC’ (<1 ppm/°C) current can be found. This process of selection of operating current was done for many years (1975-1990) in order to produce accurate references for premium 5½/6½ digit DVMs.

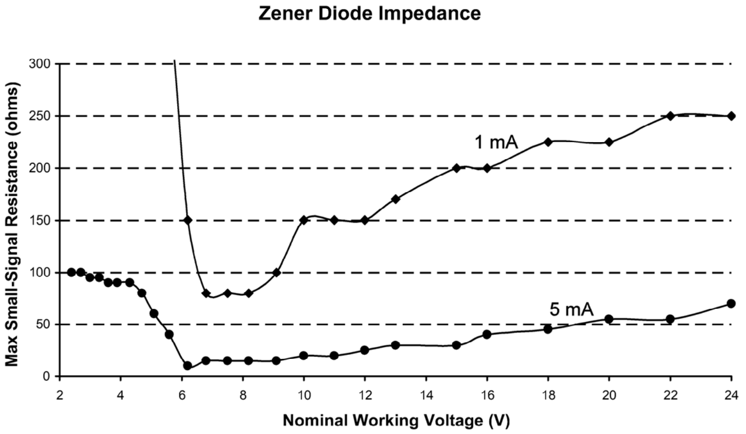
Historically low-current standard zener diodes were not well specified until the reverse current reached 5 mA. Some were partially defined at 1 mA. A new generation of low-power zeners now exists that is specified at 50 µA (BZX99). But don’t just take an ordinary zener and run it a this lower current, you need a properly specified device.

For micro-power operation, there are integrated circuits which behave in a similar but better way, rated down to 10 µA (eg LM385). These are more complicated devices and you will pay significantly more for them. Even a cheap LM385 costs $0.60. If you spend more money you can get an LT1389 band-gap reference which requires only 0.8 µA, but at $4.00.

Consider the everyday $0.04 zener. Below 5 mA it is not fully specified. If you rely on it following an exponential characteristic below 5 mA you are a novice. The characteristics below the ‘knee’ will vary horribly with temperature, horribly from batch to batch, and horribly from manufacturer to manufacturer. Always bias the zener above this knee in the curve to give the defined performance. Above the knee the zener is well behaved and has a defined slope resistance, the *dynamic resistance*. Changes in current produce changes in voltage in the usual $\Delta V = \Delta I \times R$ manner.

The zener diode is ordinarily used as a cheap *shunt regulator*. It is cheaper than a proper voltage regulator, but its performance is not wonderful. Below is real data from the Philips BZX84 series at 25°C. Notice that the dynamic resistance is very much higher at 1 mA compared to 5 mA. Also be aware that the operating voltage is specified in ±1%, ±2% or ±5% bands (selection tolerance), but only at 5 mA. If you wish to run them at 1 mA, you have to evaluate the operating voltage for yourself.

FIGURE 8.7A:



EX 8.7.1: Using data from the above graph, estimate the worst case LF ripple rejection of

a 3.3 V zener diode when fed from a 12 V supply for the case of:

a) An 8.2 kΩ resistor.
b) A 1.8 kΩ resistor.

***EX 8.7.2:** A +15 V ±5% power rail is available, but you need a well regulated supply of about 6 V for an interferometric phonon multiplier.† This device takes a varying load current between 1.0 mA and 3.3 mA, regardless of the supply voltage. There are plenty of 6.2 V zener diodes in stock. Their spec is 6.2 V ±5% at 5 mA, with a dynamic resistance of not more than 10 Ω. Ignore temperature effects on the zener.

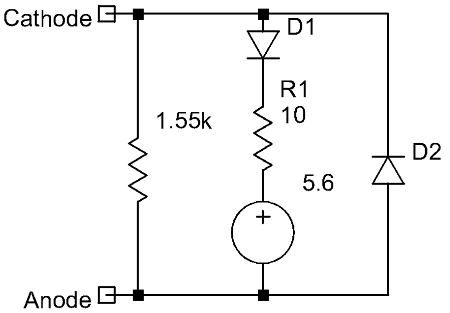
a) What resistor should be used to supply the zener from the 15 V rail?
b) What is the maximum ripple on the zener voltage when the supply is nominal (and assumed noise free)?
c) What is the ripple on the zener due to a 300 mV ptp 120 Hz ripple on the +15 V rail?
d) What is the ripple rejection in dB on the zener with respect to the 15 V rail?

EX 8.7.3: A 4.7 V zener diode is supplied with 20 mA from a good constant current source. The load connected to the zener suddenly changes from 1 mA to 15 mA. Describe the voltage seen by the load in quantitative terms. Just use the nominal data:

V_Z =4.7 V @ 5 mA; R_Z = 65 Ω; $R_{TH\ J\text{-}A}$ = 430°C/W; TC= −1.4 mV/°C

Looking further on the 6.2 V zener data sheet you find that it is specified to have a leakage current of no more than 3 µA at a reverse voltage of 4 V and a junction temperature of 25°C. This makes you think that it is ok to operate the zener at low reverse voltages, but it is not ok. The problem is that you have not been told how this current varies with temperature. If it doubles every 10°C then by the time you reach an operating ambient of 75°C, this harmless 3 µA leakage current has turned into a harmful 96 µA. There is also no evidence or guarantee that this current does double every 10°C and it is risky to assume that it does. Obviously this leakage effect gets progressively worse as you get closer to the knee.

FIGURE 8.7B:



This large-signal (5.6 V) zener diode equivalent circuit is a *training aid* not a simulation model. Unless you bias the model at more than 4 mA there will be insufficient current to make D1 conduct. (Consider D1 as perfect; no volt drop in forward conduction). R1 is the dynamic resistance and D2 gives the forward bias conditions.

Use this model until you get the idea that *you must bias the zener to get a well defined dynamic resistance*. Once you get this idea you can discard the above model.

† Invented product type.

If you take a micro-power zener and run it at 50 μA you will get an extraordinarily high and unspecified dynamic resistance. However, at the same current, one of the more complex micropower references will give a low and well defined dynamic resistance. This just follows the general rule that you can get better performance by adding more (internal) complexity and/or paying more money.

It is very important to state clearly that a zener diode is noisy. For most applications it is *essential* to decouple them with at least 10 nF; 100 nF is better, and it is not at all unusual to use 10 μF, possibly with 100 nF in parallel. The capacitor not only reduces the noise, it also reduces the source impedance; changing loads do not then produce unpleasant glitches on the zener output.

The lowest zener voltage available is 2.4 V. Below this you can use a forward biassed double diode to give about 1.4 V or a forward biased single diode to give about 0.7 V. Single and double diodes are also available that are specifically characterised for use in voltage reference applications. These give well defined voltages, compared to the inadequate data normally given for diodes.[†]

8.8 The Varicap (Varactor) diode.

The capacitance of all diodes reduces as the amount of reverse bias increases. The Varicap diode, however, is optimised and specified for this application. The maximum capacitance achievable for a varicap is around 500 pF, with the ratio of high to low capacitance for any particular type being <20:1.

Typical applications include Voltage Controlled Oscillators, electronically adjustable HF compensation networks, electronically adjustable phase/time ***skew*** adjustment schemes, and frequency multipliers.

You should know by now that you must always use the *minimum* amount of adjustment possible in any circuit. The variable element is almost certainly the least stable element; using it as the main element in your system then gives the worst possible TC, long term stability, noise, non-linearity… need I go on? Normally it is best to de-sensitise the range of the varicap by use of either a small series capacitor (less than the capacitance of the varicap) or a parallel capacitor (greater than the capacitance of the varicap).

Varicaps can have *abrupt* or *hyperabrupt* junctions, referring to the doping density gradient at the junction. The abrupt type gives a high Q and has a wide range of tuning voltage. The hyperabrupt type has a more linear voltage-to-capacitance characteristic, but with a lower Q. Abrupt junctions are therefore preferred for the lowest possible *phase noise*, whereas hyperabrupt types are preferable for a more linear control range.

The capacitance change with reverse voltage for an abrupt junction is given by:

$$\boxed{C_j = \frac{C_{j0}}{\sqrt{1+\frac{V_R}{\Psi}}}}$$

C_j is the actual junction capacitance.

C_{j0} is the zero-bias junction capacitance.

V_R is the applied reverse bias.

Ψ, {psi; pronounce like *sigh*} is the built-in junction potential.
[$\Psi \approx 0.7$ V for silicon and 1.2 V for gallium arsenide.]

[†] Philips BAS17.

This formula can also be used to estimate collector-base junction capacitance on transistors.

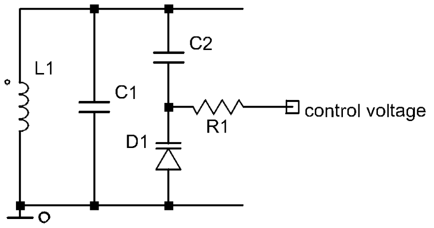
The Q of a varicap is not always easy to compare with other devices of the same type. The reason for this is that the Q varies dramatically with the amount of reverse bias, due to the change in capacitance.

$Q = \dfrac{1}{2\pi f R_S C}$. Unless the diodes are specified under exactly the same conditions it can be difficult to compare devices from different manufacturers and indeed to compare the hyperabrupt and abrupt types. The trick is to convert the Q value given by the manufacturer to an equivalent series resistance (ESR). Whilst the Q can vary by 10:1 over the operating range of a particular varicap, the equivalent series resistance will be fairly constant.

Consider a specific example. The MicroMetrics abrupt varicap MTV4030-18 is specified as having a capacitance of 10 pF with 4 V of reverse bias. It has a Q of 2400 @ 50 MHz and the same bias voltage. This gives a calculated ESR of 0.13 Ω. The hyperabrupt TV2101, from the same manufacturer, also has a 10 pF capacitance at a reverse voltage of 4 V. However, it only has a Q of 400 @ 50 MHz, giving an ESR of 0.80 Ω. In this case the direct comparison of Q factors is easy, but in general it is useful to be able to look at the relative quality of the parts by looking at the ESR.

Let's suppose you are making an oscillator that is phase-locked to exactly 100 MHz. This is a fairly typical application for a varicap. You may well have decided to make an LC oscillator of some type, trimmed by the varicap. Again this is all very standard.

FIGURE 8.8A:



In this partial circuit, just showing the resonant circuit, you see the DC blocking capacitor C2. Now C2 can be made large so that all of the varicap range is available to trim C1. It can also be made small, so that the range of the varicap diode has a lesser effect on the resonant frequency. C2 then has a dual function of DC blocking and setting the adjustment range of D1. This is fairly common in oscillators >30 MHz, since you don't need much capacitance range and the lowest capacitance varicaps available are still too high. It has the added advantage of reducing the signal swing on D1, reducing the harmonic distortion. R1 has to be large enough not to damp the Q of the resonant circuit, but not so high that the leakage current in D1 causes a changing volt drop across R1.

Let's suppose you have used a 15% tolerance inductor and are using 10% capacitors as standard. Therefore you determine that the varicap should be able to trim out those errors. Since you want the LC product to be at the fixed value necessary for the oscillation frequency of 100 MHz, you decide to design the circuit to have 25% trim range by means of the varicap. That is the cheap solution, but it may not work well in your application.

Remember that a varicap is a voltage variable capacitance. The signal therefore modulates the capacitance, distorting the signal, the degree of distortion depending on the output signal size. The distortion also depends on how much of the total capacitance is variable. The less that is variable, the less the distortion.

Maybe you don't care about harmonic distortion on the oscillator. Fine, but do you care about ***jitter***? If the varicap has twice the control range then noise on the control line will have twice the effect. Perhaps you think the phase-locked loop will take all that out. Dream on.

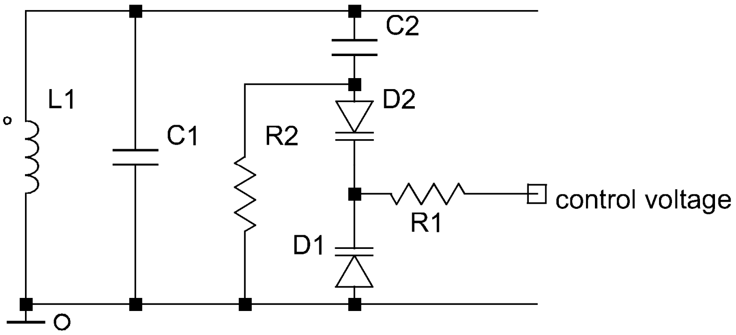
There is a wide-spread (false) belief that a phase-locked loop solves all noise and stability problems. Let's get something straight about phase-locked loops. If you have noise on the power rails to the voltage controlled oscillator (***VCO***), the loop will correct for it by varying the phase of the oscillator. You will therefore introduce phase noise onto this otherwise wonderful oscillator. You will also get unnecessary phase noise if you have too much control range in the VCO. If you are not doing a demanding design, and phase noise/jitter are not important then go right ahead, use all the range you want.

If you have a demanding application then you need to:

- reduce power supply noise by local filtering
- reduce ground noise by use of a separate ground plane
- use a tight selection tolerance and TC for the inductor and capacitor
- include a trimmer capacitor (probably cheaper than a variable inductor)
- fit a metal screen over/around the oscillator and phase detector.

Having done a coarse manual adjustment of the oscillator, you can now reduce the range of the varicap to take into account only temperature and time stability of the components in the oscillator. As a final check, heat the circuit with a hairdryer and see if it drops out of lock. (Just a crude but quick test for enough control range.) Try the same test (gently) with freezer spray. These freezer sprays can take the circuit down to –50°C, so you have to be careful not to overdo it.

FIGURE 8.8B:



If you need a big signal on the varicap (eg >1 V ptp) the distortion can be reduced by putting a pair of identical devices in series opposition. The even-harmonic distortion products will then be minimised. Make both R1 and R2 many times larger than the capacitive reactances in the circuit.

Varicaps are very difficult parts to second source since different type numbers are never really compatible. Generally you should pick a single manufacturer and keep a larger stock level to cope with product availability problems.

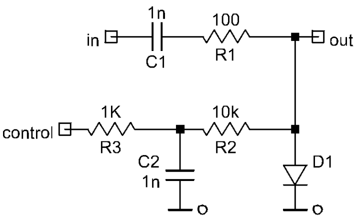
HBV (hetero-structure barrier varactor) diodes can be used to make efficient frequency multipliers (eg triplers) for operation above 100 GHz. The power conversion efficiency of such a tripler is not limited to $\frac{1}{3^2}$ as it is for schottky diode triplers.

8.9 The PIN diode

The name is sometimes written as P-I-N since it means *P*-type material, *I*ntrinsic material then *N*-type material. In catalogues they may be called *band-switching diodes*.

The PIN diode is relatively inexpensive for a basic type, let's say $0.25. They are very good as VHF switches since they have low capacitance when 'off' and low resistance when 'on'. They can also be used for variable VHF attenuators. The action generally takes place at frequencies above 30 MHz. If you want them to work at lower frequencies then you have to pay more money.

FIGURE 8.9A:



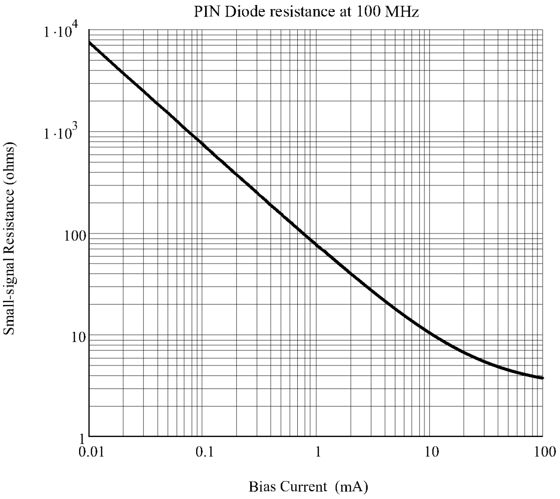
This is a simple VHF attenuator using a PIN diode. The control voltage swings from +5 V to −5 V, negative voltage giving minimum attenuation. The capacitance of the PIN diode, and ordinary diodes, reduces when the reverse bias is increased, only a few volts being necessary to get most of the benefit.

Notice that R2-C2 stops the VHF signal leaking into the control circuit and R3-C2 prevents noise from the control circuitry getting onto the main signal. This is a standard design technique. Get used to adding these simple filter components to your designs.

If D1 were an ordinary diode then it would still attenuate the signal, but its on-state resistance would be more like 26 Ω at 1 mA bias current. The PIN diode is 'magic' in that the bias current can be less than the signal current and the circuit still functions correctly. It is a VHF switch, not a DC switch, so this is possible.

Be warned that this is a typical 'text book' circuit section, in that it does not have much attenuation range and it does not have much drive capability. It is presented here to give an idea of the type of circuit the PIN diode could be used in. Manufacturer's application notes will tell you more.

FIGURE 8.9B:



The key parameter on a PIN diode is the *carrier lifetime*, τ. A rough rule is that at frequencies (in Hz) below $1/(10\pi\tau)$ the PIN diode acts like an ordinary silicon diode. At frequencies (in Hz) above 1/τ the carrier lifetime is long enough that when the reverse conduction cycle occurs, the carriers are still available for conduction. In this case the diode looks more like a resistor than a diode. The resistance is controlled by the DC current, less current making the resistance greater in a very well specified manner.

The PIN diode small-signal resistance characteristic is accurately modelled by …

$$\boxed{R_D = R_{REF} \cdot \frac{I_{REF}}{I_D} + R_{MIN}}$$

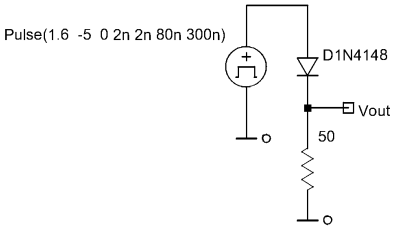
The effective small-signal RF resistance of the PIN diode curves down to a minimum value at high current. At lower currents the resistance is inversely proportional to the DC current.

Whilst PIN diodes have historically only been available for VHF and above, they are now available for use down to 100 kHz (For example, Philips BAQ800).

8.10 The Step Recovery Diode (SRD).

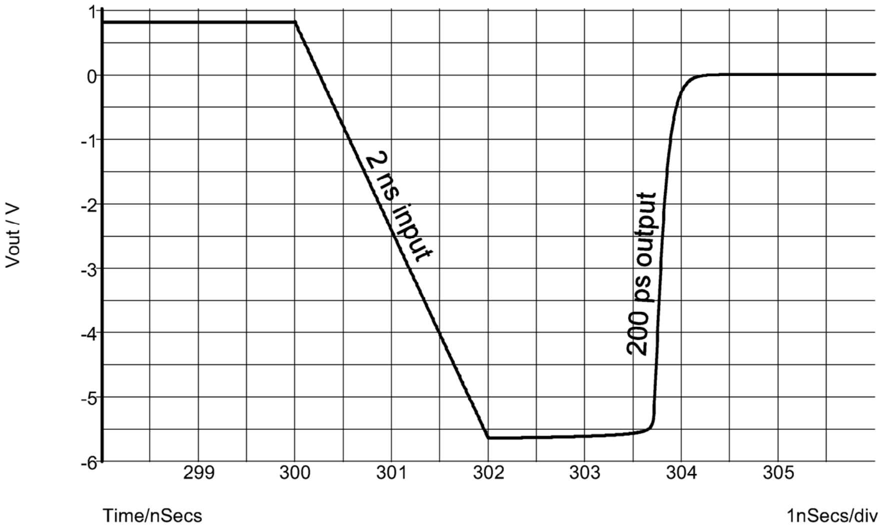
In the section on rectifiers, the simulation of the IN4002 gives an idea of the nature of the step recovery diode. The diode is first forward biassed to build up charge in the junction; it is then rapidly reverse biassed. The diode conducts some reverse current, but as soon as the minority carriers are depleted, the current stops very abruptly. This is the step in the *step recovery*. In switched-mode supplies you don't want a step recovery (also called snap recovery) because the fast transient generates EMI. (Refer back to Figure 8.5F, page 117). In the step recovery diode this fast step is exactly what you are trying to produce.

FIGURE 8.10A:



This simulation model of a fast pulse generator uses an ordinary diode, since they simulate just like a step recovery diode. Step recovery diodes are standard catalogue items for microwave component suppliers, but they are not cheap. They cost upwards of $30 each, but are readily available with edge speeds to < 40 ps. Faster edge speeds are not as readily available.

FIGURE 8.10B:



You don't get out a larger voltage pulse than you put in to this circuit. What happens is that you put in a fairly fast edge and get out a *very* fast edge.

8.11 The Tunnel Diode.

Discovered in 1957 by Esaki, the tunnel diode is sometimes referred to as the Esaki diode.[5] It is the one of the most expensive diodes mentioned here for low power devices; price guide say $100.

Rather than having a monotonically increasing bias characteristic, the voltage across a tunnel diode at first rises with increasing current, but then falls sharply before continuing to rise again. This negative slope resistance region is what makes the tunnel diode capable of amplification and oscillation at UHF frequencies. The voltage depth of the negative slope resistance region is a measure of the fast edge generation capability of a particular device.

The tunnel diode is a very tricky component to make. The production yields are very low and the costs are therefore high. Tunnel diodes have traditionally been used to produce very fast pulses (in the range 1 ns to <25 ps), but such generators have a great limitation of repetition speed. It is not unusual for the 200 ps and faster generators to have repetition rates of 100 kHz and below. Tunnel diodes are therefore only used in cases of extreme need, where nothing else can deliver the performance.

Tunnel diode generators themselves can be very delicate, being susceptible to everyday static discharges. For this reason, great care has to be taken when using these generators. It is not unusual, for example, to have to discharge a coaxial cable before connecting it to a tunnel diode pulse generator in case there was any static charge existing on the cable. The 100 pF of a 1 m length of coax can store sufficient static electricity that it can be hazardous to an ESD sensitive input or output.

With modern silicon and gallium arsenide devices, it is possible to make pulse generators with fast repetition rates and respectable risetimes. Fairly ordinary silicon devices can give risetimes around the 300 ps region. Much faster than this and you need SiGe and GaAs devices. Step recovery diodes are the way to go if you can tolerate the slow repetition rates or if you need to generate >50 V pulses.

8.12 The Gunn Diode

The Gunn diode [6] has a negative resistance region which makes it capable of producing oscillations. When correctly supported inside a mm-wave cavity, tens of milliwatts of power can be generated at mm-wave frequencies up to and beyond 100 GHz. It is the dimensions of the cavity which help define the oscillation frequency, although the device itself will only have at most an octave of usable operating frequency range. As with all microwave/mm-wave parts, Gunn diodes for specialist applications can cost >$1000, although high volume parts for automotive sensing applications can be remarkably inexpensive ($20).

IMPATT diodes can also be used for mm-wave sources. Unfortunately they are such broad-band devices that IMPATT oscillators are notoriously difficult to design and produce excessive phase noise. Imperfect design can then result in high spurious tone levels (–20 dBc) and instability of the fundamental mode.

[5] L. Esaki, 'Discovery of the Tunnel Diode', in *IEEE Transactions on Electron Devices*, ED-23, no. 7 (1976), pp. 644-647.

[6] Named after J.B.Gunn who first identified their transit-time mode of oscillation in 1963.

8.13 The Synchronous Rectifier

Whilst a silicon diode has a forward volt drop of ≈0.6 V, and a silicon schottky diode has ≈0.4 V, a synchronous rectifier can have a volt drop of < 100 mV. Depending on the current it can be even lower than that. The reason is that a synchronous rectifier is not a rectifier at all, but a switching device. It could be a bipolar transistor, a FET, a reed-relay, or a mechanically operated switch such as the commutator on a DC motor.

The low volt drop means that the synchronous rectifier can be much more efficient than a rectifier diode. The penalty is increased cost and added complexity. Any support circuitry for this switching device has to consume very little power if the overall power dissipation is to be reduced.

The term ‘synchronous’ means occurring at the same time. The active device is turned ‘on’ just when it needs to behave as a rectifier in the forward direction, and it is turned ‘off’ to resemble an off-state rectifier.

The synchronous rectifier may not be just one component; it is the switched device along with the necessary control circuitry. This is often built into switched-mode controller chips so the key thing to know is that the concept exists. It is necessary to use this type of scheme when large currents and low output voltages (< 5 V) are present, and/or where high efficiencies are required.

8.14 Shot Noise

When charge carriers randomly cross a barrier such as a p-n junction the current has a noise associated with it. This noise phenomenon was quantified in 1918 by Walter Schottky[7] resulting in this formula:

$$\boxed{I_N\,(RMS) = \sqrt{2\,e\,\Delta f\,I}}$$

The mean-squared noise current is proportional to the charge on an electron (e), the effective measurement bandwidth (Δf) and the diode current (I).

The voltage noise across a diode decreases as the current increases since the small-signal resistance decreases faster than the shot noise current increases.

The small-signal resistance is $\boxed{R_D = \dfrac{\eta V_T}{I_D} + R_B}$, R_B is the fixed ‘bulk resistance’.

At low current levels the bulk resistance is negligible compared to the current dependant resistance, making the small-signal shot noise …

$$\boxed{V_N(RMS) = \eta \cdot V_T \sqrt{\frac{2\,e\,\Delta f}{I}}}$$

… which is about 2 nV/√Hz at 100 μA.

[7] Van Der Ziel, A., 'History of Noise Research', *Advances in Electronics and Electron Physics*, 50 (1980), 351-409.

CH9: the transistor

9.1 Historical Overview

The transistor is a relatively recent invention, the first point-contact transistor being demonstrated in 1947. The name originally comes from TRANSfer + resiSTOR, which is how the gain of the device was being explained at the time. The first satisfactory junction transistor was not demonstrated until 1950.

These early transistors did not immediately replace the thermionic valves {vacuum tubes} that were the key amplifying devices of that time. A 1959 magazine article stated, "… an HF transistor for television receivers has not yet been produced commercially."

In 1967 valves and transistors were still being compared, with valves being better in several applications. The upper frequency response of 1 GHz for transistors was slower than the 10 GHz of valves. It was only around 1995 that the 10 GHz barrier for silicon transistors was broken. Silicon-germanium (SiGe) smashes this barrier by providing greater than 50 GHz performance. For amplifiers above 5 GHz bandwidth the device of choice is a HEMT (High Electron Mobility Transistor), also known as a GaAs FET. The complaints of the transistor's sensitivity to nuclear radiation and its limited power output (300 kW claimed for a valve!) are still valid today.

Valves are still being used for the output stages of high power broadcast transmitters. For example WWVB, the 60 kHz frequency reference transmitter in Colorado (USA), has valves in the output stages of its 50 kW amplifiers. It should be remembered that vacuum-electron technology is also required for kilowatt to megawatt microwave radar, and other systems, in the form of magnetrons and klystrons. Backward Wave Oscillators (BWO) are also vacuum-electron devices and these can produce power up to 1000 GHz.

When I speak of transistors, I mean *bipolar junction transistors* (BJT). A *Junction Field Effect Transistor* will be referred to as a JFET (jay′ fett) and a *Metal Oxide Semiconductor FET* will be referred to as a MOSFET (moss′ fett).

If you wanted to know all about transistors then the ideal time was the early 1960's. Transistors were new, small, thrilling and *expensive*; this last factor is important because if a device is sold for lots of money, the manufacturer can afford to spend time and money characterising it. The manufacturers were pushing them hard and lots of excellent texts were produced to explain to the existing designers how to use these brand new inventions.

If you were a designer in the 1950's then you would have been using valves. When transistors came out, changing over to transistors was difficult. Articles and books had to be written to 'convert' existing designers over to these new delicate devices. By this time, continual improvements had made valves very robust, electrically, whereas transistors would 'blow up' {fail} at the slightest overload.

If you look at books from the early 1960's they explain transistors very well, letting you know exactly what to do with them. The modern trend has been to confuse the issue by teaching hybrid-π models before learning how to do something simple like light an LED with a transistor, for example.

It is important to know something about the insides of a transistor, but in-depth knowledge of the underlying physics of semiconductors is not necessary to use the devices proficiently. After all, a transistor is a three terminal amplifying device. Consider it as such, so that when organic superconducting transistors are invented, you will realise that only bias conditions and impedances have changed!

Look at modern transistor circuits and then look at the earlier valve circuits; as far as *basics* are concerned they are the same. Bias conditions and impedances are all different, but the concepts of the ***cascode*** circuit, the ***Colpitts*** oscillator, and the ***Schmitt trigger*** are all the same. Even noise considerations are similar. [Johnson did the work on thermal noise with a valve amplifier.]

Transistors in small TO-92 plastic packages come in every permutation of emitter, base and collector (E-B-C) relative to the package that you can think of. At least with surface mount SOT23 packages, and the newer smaller ***footprint*** devices, the collector is always the single leg, with the base and emitter in fixed positions.

The bipolar transistor is the cheapest and most commonly used of the transistor types for discrete circuits, a small-signal type (BC847) costing ≈$0.03 in volume.

Designing simple switch circuits, regulators and general interface circuitry at the transistor level is not necessarily either trivial or complicated. It is really a question of spec. Making discrete transistors into accurate amplifiers (better than ±0.5%) at low frequencies or high speed, however, is always hard, and frankly nowadays a relatively obsolete skill.

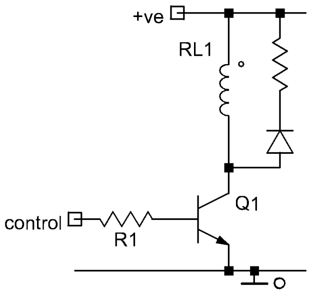
In the early days of commercial transistors (the early 1960s) the amateur electronics magazines were full of radio kits. Build your own transistor radio! With this new invention, transistorised portable radios could be built by the amateur. The adverts didn't say that this kit had 20 dB better sensitivity than another, or 5kc/s [†] more bandwidth. The adverts simply said build a "5-transistor radio" or a "7-transistor radio". Having more transistors was perceived to be better. (Even today, microprocessor manufacturers boast that their chip has 10,000 gazillion [‡] transistors in it.)

Good thinking: more transistors *can* make a better circuit. The ultimate of using more transistors is to use a packaged device such as an *opamp*.

In the circuits that follow, ***+ve*** is just a shorthand notation for positive.

9.2 Saturated Switching

FIGURE 9.2A:



If the control is guaranteed to drop to <0.4 V then there is no need to put a resistor from the base of the NPN transistor to ground or to a negative rail. A CMOS driver guarantees this; an old fashioned TTL driver does not.

This is not really a 'design task'. This is a 5 minute, back-of-an-envelope task.

EX 9.2.1: The above circuit works in an ambient from 0°C to 70°C. The load is a relay

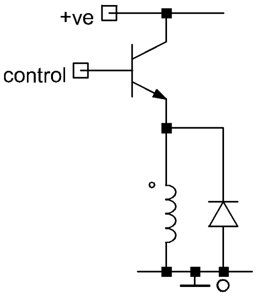
[†] 5 kc/s, is 5 kilo-cycles per second: 5 kHz in modern units.

[‡] Invented word for a large number; any real figure would be almost immediately obsolete.

coil of resistance 1 kΩ ±10%. Its must-operate voltage is 0.8 of nominal at high temperature. Its must-release voltage is 0.1 of nominal at all temperatures. The nominal operating voltage is 12 V and the power rail is 12 V ±5%. Neglect the source impedance of the control voltage and assume it has two states; +0.3 V and +11.7 V. The transistor's H_{FE} is specified as between 70 and 150 at 25°C at a V_{CE} of 5 V. Taking into account all practical considerations, what component should you use for R1?

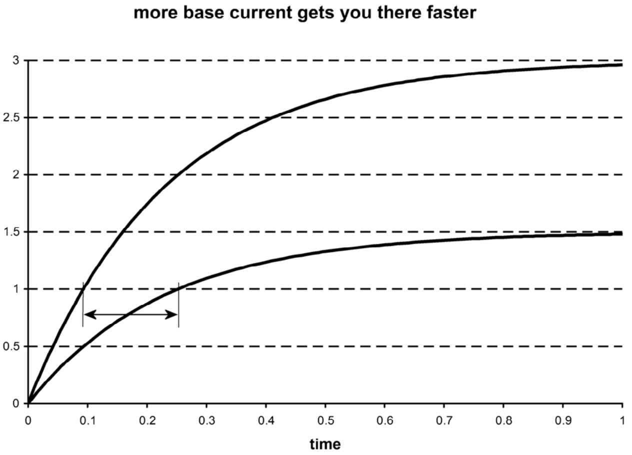
***EX 9.2.2:** Actually, your boss has just told you that you are a waste of space[†] using components so extravagantly in that previous circuit. He draws this next circuit on the back of an envelope.

FIGURE 9.2B:



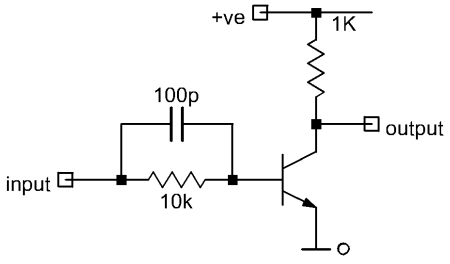
What is your advice? Be prepared to justify your answer because there are one hundred of these drivers per module and you are due to sell one hundred a week. The boss is looking to save money. He might get rid of you {dismiss} if you insist on wasting components so needlessly.

FIGURE 9.2C:



The idea is that there is an 'aiming value' of collector current given by $H_{FE} \times I_B$. This current would only occur if the collector was tied to a power rail without a collector resistor. Aiming for a higher value gets you to the limiting value, set by the collector resistor, much faster. To turn a transistor on quickly you have to give it an excess of base current.

FIGURE 9.2D:



This is the classic 'speed-up' capacitor. It is **very** important as it speeds up the turn-on by 10× for the cost of capacitor ($0.03). 100 pF is a good starting value for a small-signal circuit. (Increase the capacitance for a heavier load.) The quickest way to find the optimum capacitor value is to physically change the capacitor until the transistor switches quickly!

The turn-on time could be reduced by reducing the base drive resistor. The problem is that running the transistor with excess base current, and therefore excess base charge, makes the turn-off slower. The speed-up capacitor handles this by initially giving the

[†] term of abuse.

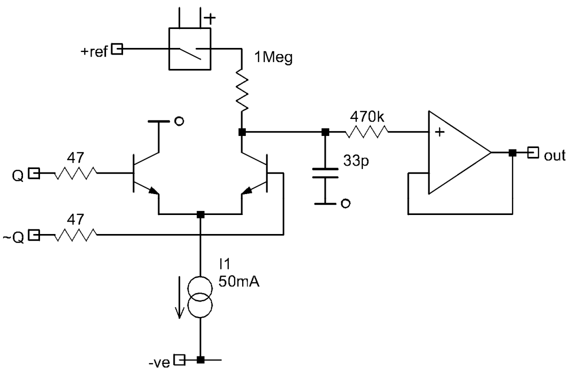
base more current, but just on the edge.

Q1 is unlikely to be an RF transistor; if it were then it might oscillate during the switch on period. However, a high voltage transistor may also oscillate during the transition; if so then put a 68 Ω resistor in series with the capacitor, a *base-stopper resistor*.

The design technique is to use the base resistor to supply the minimum necessary base current found from the collector saturation curves, if available. The current gain with $V_{CE} < 0.6\text{ V}$ is not the quoted H_{FE}, but is likely to be >10. If $H_{FE} > 100$ start with $H_{FE}/10$ as an estimate of saturated current gain. Supply more current on the edge using the capacitor, driving the base even harder initially.

What I have described here is *saturated switching*. It is slow compared to current steering for a given speed of transistor. As a rough idea, *emitter coupled switching* can be >10× faster.

FIGURE 9.2E:



This is a simplified version of a real time interval measuring circuit; it can measure time intervals from 0 ns to 2 ns with a useful resolution of better than 20 ps. The Q and ~Q inputs are from a differential ECL logic gate. [The high level is –0.7 V and the low level is −1.8 V.] The key to this circuit is that you must not try to turn the transistor on for only 20 ps, nothing would happen.

What you should do is always turn the transistor on for a minimum of say 2 ns and then start measuring time *differences*. In this application the actual timing was from 2 ns to 4 ns in order to get the 2 ns range. This is part of a more general rule: Don't make the high to low range of a measured quantity too large, or you will be making life unnecessarily hard on yourself. By this I mean the ratio of the maximum to minimum value should not be too high. Measuring from 0 ns to 2 ns gives an infinite ratio of max over min, suggesting that it is a difficult measurement at best.

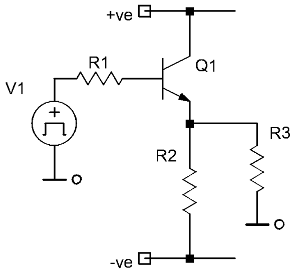
9.3 Amplification

The previous circuits were just using the transistor as an electronic on/off switch. This is the most basic application of a transistor and you should be able to use a transistor in this way without spending more than a few minutes on the component values. Let's complicate the situation and use the very simplest sort of amplification, the *emitter follower*.

The emitter follower is also called a common-collector circuit, the idea being that the collector terminal is common to both the input and output signals. Thus texts refer to the three possible configurations as *common-emitter*, *common-collector* and *common- base*.

In an emitter follower the transistor acts as an *impedance buffer*, with a voltage gain of slightly less than 1, and a current gain in the region of 40 to 400. Note that text books use the term *gain* to mean overall *power gain*. Thus a transformer does not have gain because although the voltage can be 'stepped-up', the power is not increased. Because of the current gain, an emitter follower does have power gain.

FIGURE 9.3A:



In this simplified circuit, R1 represents the source resistance and R3 represents the load resistance. The whole point of the circuit is that R1 >> R3. If R1, the *source*, were to be directly connected to R3, the *load*, there would be a huge loss of signal.

EX 9.3.1: R1=1K; R2= 1K; R3= 200 Ω; The +ve rail is +12 V; the –ve rail is –12 V. Assume that the bulk emitter resistance of the transistor is 3 Ω and that the base-emitter junction has a TC of –2 mV/°C.

a) What would the voltage attenuation be if Q1 and R2 were left out, the load being directly connected to the source?
b) Estimate the voltage attenuation for small low-frequency signals when V1 has a DC level of 0.7 V. Use a transistor with an H_{FE} of 100 and neglect tolerances. Assume room temperature operation of the silicon. Neglect self-heating

In that case, the gain of the overall stage was dominated by the input resistance of the emitter follower, not by the output resistance. The current in the transistor can therefore be allowed to change quite a lot, provided that the current gain does not change too rapidly with collector current. In practice, transistor current gain characteristics are fairly constant over at least a decade of collector current variation, and often they are fairly constant over many decades.

Whilst complex transistor models are presented in early courses, you should now use considerably simplified models instead. More complex situations can just be simulated using SPICE, or similar software, which is freely available on the Internet. The output resistance of an emitter follower [or more generally the impedance looking back into an emitter] is composed of three parts:

- the fixed *bulk emitter resistance.*
- the resistance supplying the base divided by the h_{fe} of the transistor
- the reciprocal of the transconductance, $\left[1/G_M\right]$

The bulk emitter resistance can be somewhere between a few parts of an ohm [for a large transistor] to several tens of ohms [for a very small transistor]. 1 Ω – 10 Ω is a typical range. This bulk resistance is relatively constant with collector current and so does not have much impact on linearity. The input source-resistor {Thévenin equivalent resistance of the circuit connected to the base} divided by the current gain of the transistor is also a relatively constant term, since current gain is often high over several decades of collector current.

The reciprocal of the transconductance is the nasty one. At 1 mA and room temperature the emitter resistance is around 26 Ω, but this drops to 2.6 Ω at 10 mA. Fortunately, this source of non-linearity is very predictable and repeatable. The solution to the problem is to maintain the collector current at a high value. That way the non-linearity due to the transconductance variation is minimised.

EX 9.3.2: Use the same circuit and values as before, but this time, take self-heating into account. The $\theta_{J\text{-}A}$ [thermal resistance, junction to ambient] of the transistor is 300°C/W. The input steps from +0.7 V to +1.7 V. Estimate what happens at the output.

That question and its answer are not something normally found in text books. Welcome to the real world! Even the most trivial of problems involving analog components can have hidden depths. In case you didn't realise, SPICE (3F5) does not model this thermal problem. SPICE will model the temperature of individual transistors, but will not show the effects of self-heating in an individual device.

As an experiment on an SOT23 BC847 transistor, I applied a 40 mW power change and found a slow thermal response of 4 mV, settling in around 50 ms. This was in addition to the faster time constant effect. This slow time constant settling will be changed by the amount of copper the transistor is connected to on the PCB.

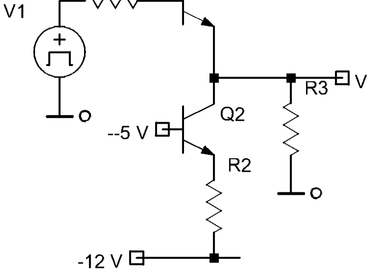
The thermal problem applies to sinusoidal signals as well, although it is more difficult to visibly see the effect on a scope. Also, the effect only occurs at lower frequencies on AC signals. Just how low is 'low' depends on the transistor size and how it is cooled. This thermal problem is easier to investigate in the time ***domain*** because you can directly measure its effect. In the frequency domain you would see the effect as a reduction in amplitude and even-harmonic distortion.

To solve this thermal problem, use more transistors. Actually the previous design was relatively easy because the load was quite high (200 Ω). Driving a 50 Ω load would have a much stronger effect on the current change for the same voltage change. Therefore the power change would be worse. Minimising the voltage across the transistor also minimises the power change.

The 'tail resistor' R3 is not helping in this circuit either. As the output voltage changes, the current in R3 changes as well. Using a larger supply voltage, and a larger resistor, would reduce the current change, but this method is not practical. The circuit is taking 12 mA from a 12 V rail, 144 mW. A 120 V rail would then require 1.44 W! This is all very unsatisfactory.

Now you could buy a two-terminal current source. These are JFETs with the gate shorted to the source. They have poor tolerances on their currents [±20%] and are quite expensive (say $0.50). They also may not be available all the way up to 12 mA. In any case, it is cheaper and more accurate to use a simple NPN transistor as the current source.

FIGURE 9.3B:



This is a simplified circuit. For good performance above 10 MHz you may need a base stopper resistor on Q2 and/or a ferrite bead in series with Q2C. Set R2 to give the same 12 mA bias current as before. Now the load is mostly only R3, Q2 acting as a current source.

EX 9.3.3: Derive an approximate expression for the power in Q1 as a function of output voltage, using the symbols: I = current in Q2, V = output voltage, R =load resistor R3, VC = power rail.

Use the above result to plot how the power changes with signal as you change the power rail. This thermal effect can be reduced, by careful choice of power supply and current source, but not eliminated. Once the error is low, a fixed correction can be added somewhere else that compensates for the error. That will be quite effective, since the thermal time-constant of the transistor will be reasonably consistent from batch to batch.

If these solutions don't give you the required accuracy then you need to "add more transistors". The ultimate in adding more transistors is to 'wrap an opamp around' the emitter follower to improve its performance at DC and LF. Thus if you are making an impedance buffer that requires accuracy of better than a percent, the first thing you should do is see if an opamp or integrated circuit buffer will do the job. Although an emitter follower looks simple and cheap, the extra bits you need to wrap around it for accurate performance make the use of a pure opamp solution cheaper.

I believe I have now shown, quite convincingly, that the emitter follower has a gain that is less than one under resistive loading conditions. Things change, however, when there is a small capacitive load. How small is *small* depends on the size of the transistor, as well as its characteristics. You should be well aware that the current gain of a transistor falls off {reduces} at high frequency.

EX 9.3.4: A transistor has a low frequency current gain of 80 and a current gain-bandwidth product (f_t) of 1 GHz. Where does the current gain start to roll off?

This roll-off frequency is of considerable importance and is known as the β cut-off frequency. β is the symbol for current gain in one of the small-signal models of a transistor. β and h_{fe} can be used interchangeably for small-signal current gain if no teachers are watching!

Above the β cut-off frequency, the current gain falls at 20 dB/decade. Looking at the output impedance of an emitter follower it could therefore be concluded that the transistor had an inductor in series with its output. Regardless of how you choose to model this effect, the fact remains that a small capacitive load on the output of an emitter follower will give a voltage gain slightly greater than 1 [say up to 1.2×] for many high frequency transistors. This simulates very nicely on SPICE.

SPICE also predicts that if the capacitance is increased sufficiently, the peaking will then be reduced, but the resulting low bandwidth would be worse than using a lower speed transistor. This region of stability for capacitive load is well known for voltage regulators, but the effect there is one of an overall loop stability, rather than the stability of an individual transistor.

The peaking effect is due to the input impedance of the emitter follower becoming negative. This is easily seen on a simulation of an emitter follower because the voltage is larger *after* the source resistor. Negative input resistance means the possibility of oscillation and it is found in practice that most transistors with an f_t ≥ 1 GHz will oscillate unless they have a resistor physically right next to the base, a *base stopper resistor*.

In order to give a definite rule, a base stopper resistor should not be more than the length of the resistor body away from the transistor base lead.

You will find that some transistors are more prone to spurious oscillations than others and it is not something that can be seen from the data sheet. Even changing manufacturers of the same type of transistor can sometimes cause a previously working design to start oscillating. The only thing you can do about this is to keep a careful control on what parts are fitted and to monitor the circuits periodically to ensure that they are not on the verge of oscillation.

The technique of adding resistance in series with the base of a transistor applies to all connection modes; ie common-emitter, common-base and emitter-follower. This is one of those elementary and yet vital parts of any design. If your design use transistors that ‘like to oscillate’ then it is essential to use base stopper resistors. However, this is not to say that one should be over-cautious and fit base-stoppers on every transistor, just to be safe. Know your transistors and use base stopper resistors only where necessary.

Some will argue with me on this point and use base stoppers everywhere to minimise design time. I know my transistors. A BC847 is a nice, well-behaved low frequency transistor. It *never* goes into parasitic oscillation and from my point of view it would be stupid to go putting base stopper resistors in circuits using it. I use a BC847 for every low power, low frequency position in my circuits, so I could end up with dozens of extra resistors if I followed the ‘safe’ path. You make up your own mind when you have had a bit of experience of design.

The value {resistance} of the base stopper resistor is again not something that you can calculate. It needs to be determined by experiment and will be in the range from 10 Ω to 100 Ω. If it is larger than 150 Ω then there is something else going on with the circuit. You could decide to ‘play safe’ and always fit 150 Ω resistors to your GHz transistors. Unfortunately all this would achieve would be to limit the bandwidth of the circuit unnecessarily. There will be an optimum value for each transistor position in your circuit. As a starting point try 47 Ω.

The reason why manufacturers don’t make RF transistors unconditionally stable by including extra base resistance within the transistor is that RF circuits typically have 50 Ω source impedances anyway, so the best RF circuit response is achieved by using the intrinsic resistance of the circuit to stabilise the transistor.

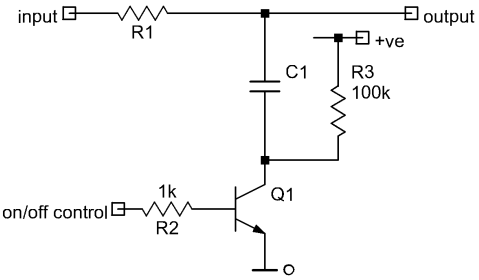
If you are working on someone else’s design, then please try changing a value here and there and see what happens. Take a stopper resistor out and see if you can reproduce the oscillations or rings. The only caution I would give you is to put it back when you are done. Very often a component is put in to fix one specific problem. It is all too easy to remove a component that is apparently not doing anything, only to find out later that it does something only in one specific bizarre and unusual condition. I have had switched-gain amplifiers that only oscillated on the most sensitive range, and only when the output was inverted, and the DC level was shifted, *and* the amplifier was cold. It would be very easy to miss that set of conditions on a short evaluation.

If you are going to permanently remove parts from an existing design, you have to evaluate the change *extensively* and that probably is not justified on financial grounds.

9.4 RF Switches

The majority of GHz transistors are not specified to any acceptable degree by the manufacturers. Often you only get typical figures for f_t rather than minimum figures, an unpleasant position to be in as a professional designer. For GHz transistors you sometimes don't even get given an h_{fe} figure to work with. The transistor may well oscillate in the h-parameter test circuit, so you get given a bunch of S-parameters instead. This is one of the reasons why discrete component >100 MHz designs are so difficult to develop and to manufacture.

FIGURE 9.4A:



This is a neat little filter / band-limit sub-circuit. It is the sort of thing that you work out and keep using over and over again. Unlike the PIN diode switch, this one has an on/off action only. The control voltage swings from –1.5 V to +5 V.

When Q1 is off, R3 defines the collector-emitter voltage. Biasing Q1 collector several volts positive gives less capacitive loading on the output, and therefore less unwanted attenuation. It also stops Q1 being reverse biased by the signal swing on its collector. R3 makes all the difference between a good switch and a poor one. R3 only has to source the off-state leakage current in the transistor. It has no significant role in terms of forward biasing the transistor.

There are two spec points to consider. The AC coupled peak signal adds to the positive bias point on R3; this must not exceed the V_{CBO} of the transistor. Also, the positive bias point for R3 must exceed the negative peak of the AC coupled signal. These two requirements are essential to ensure that the transistor is not conducting when the switch is supposed to be in the off state.

Notice that the control voltage has to go negative; this is important. On ordinary circuits you can guarantee that if the base-emitter voltage is below 0.4 V the transistor is *off*. A positive bias of any amount is not acceptable for this circuit. I recommend a reverse bias voltage of between 1 and 2 volts. Above 2 V is not a good idea as some RF transistors can only tolerate 2 V reverse base-emitter voltages.

When the transistor is conducting it behaves like a PIN diode. As such it conducts better at higher frequencies, up to some limit associated with f_t and package parasitics. Certainly at 1 MHz the base current requirement is higher than at 20 MHz. Run it at a base current of not less than 0.1 of the peak signal current in order to get a low distortion switch. This is a 'cheap and cheerful' circuit which is bound to introduce distortions. See if it is good enough for your applications by experiment.

The attenuation you can get depends on R1 and the transistor. A low capacitance transistor is needed if you either have a high value of R1 or if you wish low off-state attenuation. Q1 needs to be found by selecting a transistor with a suitably low capacitance, and this transistor then has to be tried to see how it performs. This is an 'undocumented' feature of the transistor so don't swap manufacturers of transistors without rechecking thoroughly.

On a circuit like this I would prefer a 30% margin on the measured value. But if you have < 15% margin on a measured value, compared with what you actually need, then you are *going* to get caught out. When the initial margin is < 15%, it is almost a statistical certainty that the circuit will fail to meet spec when making hundreds of units.

You can use a transistor to switch a DC signal to ground. In this case you must ensure that the signal being switched on the collector is always at least a few hundred millivolts more positive than the emitter (for an NPN transistor). The base current needed is not the peak collector current divided by h_{fe}. Remember that h_{fe} is usually specified at a collector-emitter voltage of around 5V. In order to get a good switch you need to turn the transistor on *hard*. That means more base current. As a starting point, if there is no other data available, set the base current to the lesser of $\frac{\hat{I}_C}{10}$ or $\frac{10 \times \hat{I}_C}{h_{fe}}$.

A simple test is to measure the collector-emitter saturation voltage. If it is greater than 0.4 V, the transistor is not being turned on very hard, or is being run too close to its maximum collector current. The saturated collector-emitter voltage should be made less than 0.2 V in order to give an estimated collector-emitter voltage TC of less than 0.3 mV/°C (negative). This saturated switch technique gives a low parts cost, but tends to increase the design cost, as you then have to characterise the TC for yourself. Thus such switches are often done with FETs, although the component cost goes up by a factor of around 5× to 10×.

9.5 Data Sheets

Transistor manufacturers give you data sheets on their components and from these you are supposed to be able to work out what their behaviour will be. You want to know, for example, how much voltage can be applied across the transistor before it blows up and you get given V_{CEO}. This is one of the most useless specs that anybody ever dreamed up. It is the maximum voltage that can be put across the collector-emitter terminals if the base is left open-circuit. It is a bizarre spec because the collector-base leakage current then flows into the base and gets multiplied by the H_{FE}, giving a significantly higher collector-emitter current.

So what! The base of a transistor is never left open-circuit and the test is useless at telling you what you need to know, which is how much voltage can be put across the transistor. It is certainly not as low as V_{CEO}. The correct spec to look at is V_{CBO}, the collector-base breakdown voltage. Now this is something you must not exceed; it is the absolute maximum working voltage.

There is a maximum allowable voltage across the transistor and there is a maximum allowable collector current, but they are never allowed at the same time. The maximum collector current and maximum collector-emitter voltage are *mutually exclusive* parameters. This is a favourite trick in “banner specs” on data sheets. Another favourite trick is to quote typical figures in the banner specs, meaning that you have to be careful to look at the detailed figures before deciding that a particular component is suitable for your application.

If a transistor is standing-off voltage and drawing current then it is necessarily getting hot. This means there is a voltage limit, a current limit and a power limit. But transistors have another nasty hidden characteristic called *second breakdown*.[1] You don’t

[1] H.A. Schafft, 'Second Breakdown - A Comprehensive Review', in *Proceedings of the IEEE*, 55, no. 8 (1967), pp. 1272-1288.

hear much about it nowadays because designers use power MOSFETs for switching applications; power MOSFETs do not suffer from second breakdown.

Second breakdown is a localised avalanche effect within the collector-base junction and it is usually destructive. Second breakdown limits are found on bipolar power transistor *safe operating area* [SOAR] curves.[2] You will not find this data on small-signal transistor data sheets, however. Small-signal transistors also suffer from exactly the same effect, but the manufacturers do not bother to measure or specify the characteristic.

When second breakdown is deliberately used to produce rapid (a few nanoseconds) current pulses of tens of amps, or greater than 100 V pulses with nanosecond falltimes, the name is changed to *avalanche mode operation*.[3] The transistor needs to be specifically designed for this mode or it will fail after an unspecified limited number of operations, which could be 1. Consider the ZETEX FMMT415 SOT-23 transistor. It is specified for $> 4 \times 10^{11}$ operations delivering 60 A 20 ns pulses. That is pretty spectacular performance from a surface mount package normally used for currents in the region of 50 mA to 200 mA!

What other information *don't* you get on data sheets? Well, that varies from device to device and from manufacturer to manufacturer. On some data sheets you get a well characterised device. You get curves of H_{FE} against temperature, saturation voltage at various levels of collector current to base current ratio, f_t variation with collector current, f_t variation with collector-emitter voltage and minimum specs for f_t and h_{fe}. Unfortunately you often have data sheets that don't give you this information.

The safest thing to do is to only use a transistor that is fully specified. The trouble is that the fully specified transistors may not do what you want. For transistors with f_t >700 MHz it is not at all unusual to have H_{FE} unspecified, h_{fe} given as typical or not given at all, no variation of h_{fe} with temperature given at all, and common-emitter S-parameters given instead of h-parameters.

Since all manufacturers seem to do this, the only alternative is to estimate the variations based on specified device data and try to make up your own model for the missing data. You could use MILSPEC parts, or have the parts characterised for your application, but that would add considerable additional cost. Also the very latest newest fastest devices will not be fully characterised. If you need to stay on the leading edge, using the latest hetero-junction 390 GHz f_t devices,[†] they will not be well characterised, and you will be taking a risk that the next batch of parts will not work as well as the ones you are currently testing.

If you want to run at low temperatures, say below 0°C, be aware that H_{FE} drops significantly with decreasing temperature. The problem is that many semiconductor data sheets start at 0°C and run to 125°C. Data below 0°C is just not given. The Infineon BC847 data sheet, for example, does give H_{FE} data with temperature and this shows that one cannot apply a simple rule for low temperature operation. The gain loss varies with collector current and the data given is only typical. The gain loss is significantly larger at lower currents, but at any defined operating current in your own system you might be

[2] 'Data Sheet: Transistor Safe Operating Area (SOAR)', *Power Bipolar Transistors* (Phillips Semiconductor, 1998), SC06.

[3] N. Chadderton, 'The ZTX415 Avalanche Mode Transistor', *Application Note 8, Issue 2* (Zetex, Jan 1996).

[†] Such devices do not (yet) exist!

able to estimate the possible gain loss at some low temperature by prior experience. If no other data is available, assume a low temperature (below 0°C) loss of current gain of between 30% and 40% relative to the data sheet 25°C values.

The converse is also true. Current gain increases with increasing temperature. If you run silicon at 70°C, and above, then you will get significantly more current gain. Again this is dependant on collector current, but the current gain could increase by 30% above 100°C. If you need this extra current gain then don’t rely on 30%, rely on say 10% increase for safety. The reason to use the 30% figure is when estimating the TC of the overall circuit. A bigger change of gain is more able to produce a larger gain, offset or linearity change.

Unfortunately f_t drops with temperature, to an undocumented but significant extent, so LF bipolar circuits work better when hot and VHF bipolar circuits work less well when hot. Also parasitic oscillations are more likely to occur when the active devices are cold. Hence to test an amplifier for stability {non-oscillation} it is essential to give it a quick burst of freezer spray as a very minimal test.

9.6 Differential Pair Amplifiers

For microwave/mm-wave amplifiers there is typically no need for performance below a few tens of megahertz. In this case single-transistor amplifiers can be made with AC coupling at both the input and the output. A grounded emitter or grounded source configuration is ideal as it gives the maximum possible gain.

Other than this type of use, the more usual application of discrete transistors nowadays is for buffering and interface circuitry to the main stages. In digital electronics this is referred to as “glue logic”, the odd bits and pieces necessary to join the system together. The analog system is not quite the same because there may be a single front-end stage made in discrete components to give the ultimate in low-noise performance, or the ultimate in input voltage range, or some such operation that is infeasible using an integrated solution.

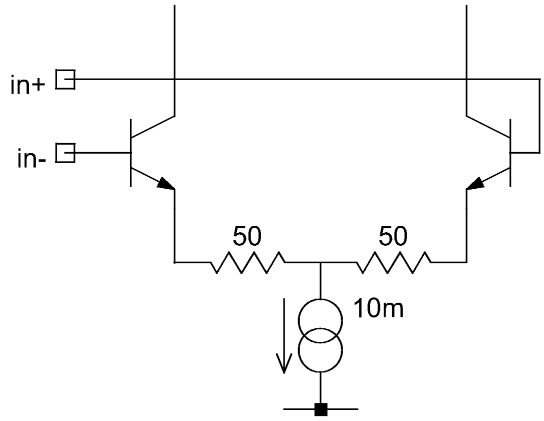
Always, always, always, look for an integrated solution first. You are making an amplifier: can you use an opamp? If not, can you use a monolithic amplifier of some other type? Once you get into this discrete component design the timescales become remarkably unpredictable. If you fail to find a suitable integrated solution then your best bet is to get hold of text books written prior to 1980. The art of designing discrete component accurate amplifiers is so seldom practiced now that modern authors may not have as much expertise as their predecessors.

If an opamp really won’t do, then a differential pair is a good place to start; it is balanced with respect to the power rails, giving better power supply rejection and therefore less ‘noise’, will work down to DC, and is inherently symmetrical so that even-harmonic distortion should be minimised.

For operation down to DC, the two base-emitter junctions in a differential pair amplifier balance each other, minimising the temperature related drift problems. Drift is further minimised by using a monolithic matched pair. –2 mV/°C is the often quoted TC of the base-emitter junction of a transistor. In practice it could be in the range 1 mV/°C to 4 mV/°C.

The emitter-coupled differential amplifier originated in 1956,[4] but this design was merely a translation of the previous valve circuitry into a transistorised form. You may wish to use differential pairs for high power stages or inside a monolithic device. In any case it is important that you know about their linearity.

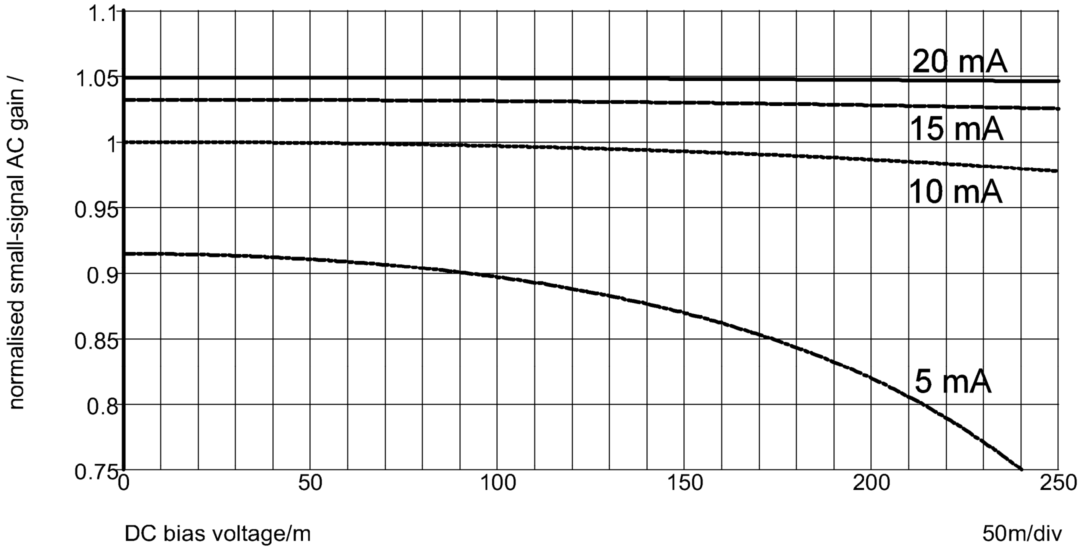
FIGURE 9.6A:



Lots of texts mention the linearity of differential pair amplifiers without emitter resistors. Simple equations cover this configuration. When emitter resistors are added the maths is unpleasant.

Let's take a design situation where the signal into the differential pair is up to ±250 mV and the emitter resistors are 50 Ω each in order to provide the necessary gain.

FIGURE 9.6B:



These curves are for emitter tail currents of 5 mA, 10 mA, 15 mA and 20 mA. They cannot be plotted directly on a SPICE simulator as they are small-signal gains done at stepped values of bias voltage. This plot needs to be done as a simulation script.

The circuit is symmetric about zero, so there is no need to plot both halves of the transfer curve. Notice that as the tail current increases, the gain of the stage increases. I have set the tail currents in multiples of the simplistically calculated 5 mA tail current deliberately. 5 mA is the minimum, and the full scale (***FS***) gain loss [relative to its maximum at the midpoint] is approximately 20%. For 10 mA the FS gain loss is approximately 2%. For 15 mA the FS gain loss is approximately 1%. For 20 mA the FS gain loss is approximately 0.3%. What is more interesting, and useful, is that these curves scale well. If you work in terms of multiples of the FS minimum current you will get similar non-linearity figures regardless of the emitter resistors used.

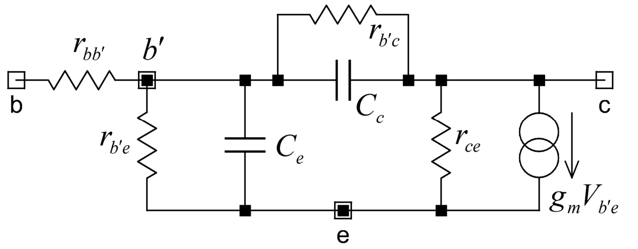
[4] D.W. Slaughter, 'The Emitter-Coupled Differential Amplifier', in *IRE Transactions on Circuit Theory*, CT-3 (March 1958), pp. 51-53.

9.7 The Hybrid-π Model

I do not expect you to be analysing circuits in any detail using the hybrid-π model, but you should at least know how to get the values of the model from the h-parameters given in data sheets. In reality if you have to model something then you should be running it up on a SPICE simulator. The SPICE model has 40 parameters to play with and it *can* model the transistor's behaviour very well indeed. SPICE models can also model the transistor completely stupidly if the parameters have been given or entered incorrectly.

Often you are given a typical SPICE model and you need to create your own max or min models from this in order to estimate some sort of production spread. You would want to have representative values of IS, BF, VAF, RB, RE, TF, CJC, CJE as a very minimum, and the max/min variation should give at least ±15% variation on these parameters in the absence of any other data.

FIGURE 9.7A:



This model is more representative of a physical transistor than others. The only problem is that the element values used are not given in manufacturer's data. They can be calculated from common-emitter h-parameters as follows:

$$V_T = \frac{kT}{e}; \qquad g_m = \frac{I_C}{V_T}; \qquad C_e = \frac{g_m}{2\pi f_t}$$

$$r_{bb'} = h_{ie} - \frac{h_{fe}}{g_m}; \quad r_{b'e} = \frac{h_{fe}}{g_m}; \qquad r_{b'c} = \frac{h_{fe}}{h_{re}} \cdot \frac{1}{g_m}; \quad r_{ce} \approx \frac{1}{h_{oe} - g_m h_{re}};$$

h-parameters have two different notations, the letter system used above and a numerical system very much like that used for *S*-parameters. If using the common-emitter *h*-parameters, there is an *e*-suffix. The equivalence between the two systems is:

$$h_{ie} \equiv h_{11e} \quad h_{fe} \equiv h_{21e}; \quad h_{re} \equiv h_{12e}; \quad h_{oe} \equiv h_{22e}$$

Taking $V_T = 26\,\text{mV}$, the typical hybrid-π values for an Infineon BC847 are:

	$I_C =$ 100 μA	$I_C =$ 1 mA	$I_C =$ 10 mA
$r_{bb'}$	11 kΩ	490 Ω	100 Ω
$r_{b'e}$	53 kΩ	8.2 kΩ	980 Ω
$r_{b'c}$	31 MΩ	28 MΩ	6.4 MΩ
r_{ce}	250 kΩ	100 kΩ	9.7 kΩ

The model values vary considerably with bias current, so for linear operation minimise the ratio max/min collector current. Specifically, if you have a signal with peak current excursions of ±2mA, do not bias the transistor at 2.01 mA. This bias level would give a max collector current of 4.1 mA and a min collector current of 0.01 mA, a 401:1 ratio.

If you bias the transistor at 4 mA, the ratio is then be 3:1; a much more reasonable proposition. The higher the bias current, the lower the ratio max/min collector currents. However, a higher bias current also means more power dissipation and can mean that the particular transistor being used will not be run at its optimum operating current. You have to study the data sheet to see the maximum collector current beyond which the h_{fe} and/or f_t start to drop off.

Thus in every application you must first decide how much power you can afford to use, then see how much bias current you want to use, then pick a transistor that can handle that much current. One has to get a feel for the circuit and try different amounts of bias current to see which gives an acceptable performance in terms of non-linearity or perhaps *intermodulation distortion*.

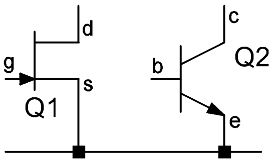
Large signal performance of amplifier stages is the most difficult area of design in terms of the available models and computational techniques. Self-heating effects within the semiconductors will give additional problems which SPICE will not model at all. Furthermore, there will also be thermal time-constants associated with coupling between semiconductors on a PCB, giving still more problems. These detailed design problems make it worthwhile to put an opamp feedback loop around a power output stage. The feedback is then used to reduce these relatively slow thermal effects.

9.8 The JFET

It may surprise you to find out that the FET is in fact older than the bipolar junction transistor in terms of its initial concept. Patents were filed by Lilienfeld in 1925 and 1928 for FET devices, but these were not taken up commercially. Shockley published the theory of the FET in 1952, but the resulting devices were not commercially available until the early 1960's, some 8-10 years after the introduction of bipolar transistors.

A JFET is a Junction FET. There is a P-N junction to the conductive channel and this junction is designed to be either unbiased or reverse biassed. It is possible for the gate to be situated symmetrically in the channel and in this case the drain and source terminals are interchangeable. For this reason the gate connection on the symbol is sometimes shown as being half way between the drain and source terminals. Otherwise the symbol is drawn showing the gate terminal as being closer to the source.

FIGURE 9.8A:



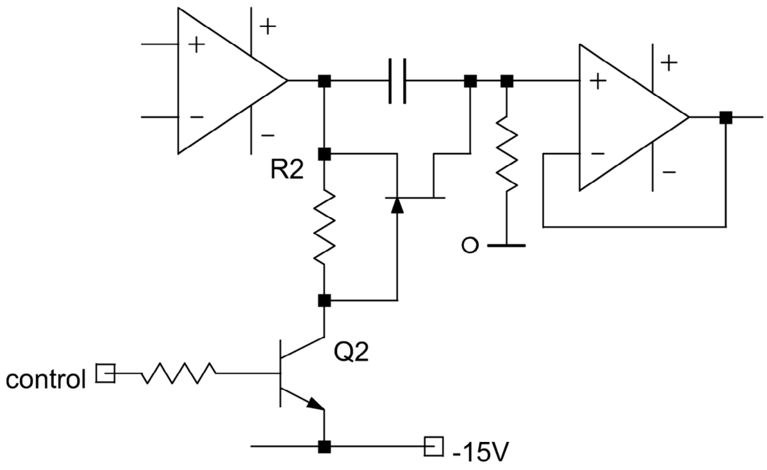
In this comparison with an NPN transistor, the phase of the output is the same. The difference is that for the *N*-channel JFET, the gate drive is a voltage which goes from 0V to perhaps –10 V, whereas for a silicon transistor the base drive is a current which develops a voltage of up to only 0.7 V across the base-emitter junction.

The (*N*-channel) JFET has a normally conducting channel of *N*-type semiconductor material. The action of the reverse biassed junction is to deplete the charge carriers from the channel thereby stopping the conduction. The JFET is therefore known as a *depletion mode* device. The *N*-channel JFET is fully ON when the gate is at 0 V and a negative gate voltage (typically between 1 V and 10V) turns it off. It is unusual, but not destructive, to forward bias the gate by a few milliamps.

There are four types of application for a JFET: an amplifier, a variable resistance [variable gain], a current source (or sink) and a switch.

Consider the JFET as a switch. To turn the *N*-channel JFET ON you make the gate-source voltage zero. This is often done by connecting a resistor between the gate and source. Because the gate leakage current is very small [pA] the resistor is usually in the range of 33 kΩ to 10 MΩ. The lower value resistance is needed when the source is moving rapidly.

FIGURE 9.8B:



Here is an example application. The JFET spec $V_{GS[OFF]}$ will probably have a value somewhere between –1 V and –10 V. To turn the device OFF you need to drive the gate by at least this much relative to the most negative of the drain and source terminals. In this example, if $V_{GS[OFF]}$ is –8 V then you cannot have the amplifier output swinging to more than –6 V or the JFET may start to turn on. R2 has to drive the capacitance of Q2 and any stray capacitance on this collector node. If R2 is too large, then when the JFET is supposed to be on, the capacitance will make the gate voltage lag behind a rapidly moving source voltage and increase the on-resistance.

That was an example circuit and not necessarily the best way of implementing a switch function. The first type of switch to consider for a new design is a packaged CMOS switch or MUX. This will be considerably cheaper than the example circuit. If this can't take the voltage, or the frequency, or the resistance is too high, then you might consider high voltage CMOS ICs.

A safe solution is often to simply use a reed relay; the on-state resistance is below 100 mΩ and the capacitance is below 5 pF. Thus for low switching rates (say <1 Hz) a reed relay can provide a more accurate and simpler solution.

When using a JFET as a switch you have no need of JFET equations or any such thing. All you need are:

- ✓ $V_{GS[OFF]}$
- ✓ $R_{DS[ON]}$
- ✓ C_{DG}

This data isn't always given by the manufacturers. If they have decided that the device is a switch then you are given this data. If they have decided that the device is a linear amplifier then they specify the device differently. It is inevitable that you will want to use a 'linear amplifier' JFET as a switch. The reason is that a device optimised as a VHF amplifier will have lower capacitances than an LF switch.

A linear amplifier JFET will be specified in terms of I_{DSS} and g_{fs}. I_{DSS} is the drain current when the drain-source voltage is held at more than a few volts by the external circuitry and the gate-source voltage is zero. This is the maximum operating current of the device. You must bias the JFET so that a worst case device (one with the lowest value of I_{DSS}) does not need to run at a higher current.

The device will probably be characterised at a specific drain current. Its forward gain characteristic is a voltage to current transfer, a *transconductance*. Hence g_{fs} is the forward transconductance when the device is in the common source configuration.

Think of an ideal JFET as a ×1 voltage buffer with very high input impedance. Now put a resistor in series with the source of value $1/g_{fs}$. This is an easy model to think with. For bipolar transistors, the output resistance of an emitter follower was a complicated sum of the bulk resistance, the base resistance referred to the output by the current gain, and an amount due to the collector current. The JFET output resistance does not have a significant contribution due to the impedance feeding the gate because the JFET's current gain is so high. Hence you just have a curve of g_{fs} against drain current.

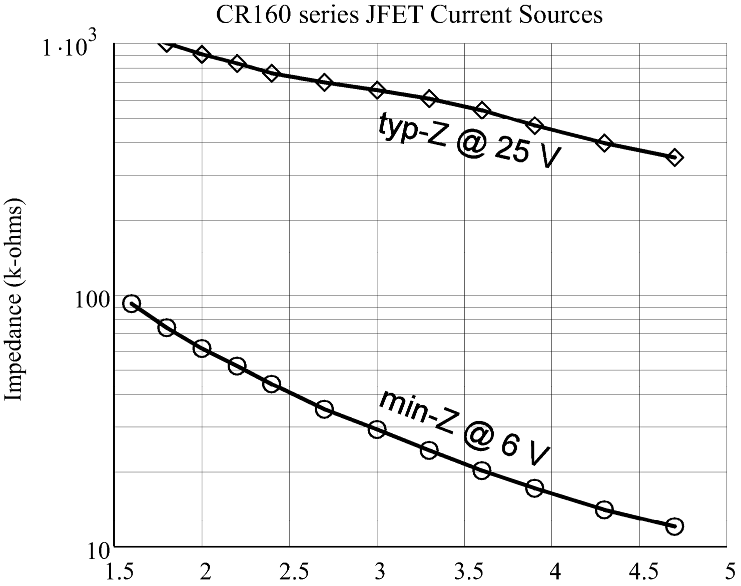
The critical factor is that the gain of a JFET is 40× lower than that of a bipolar transistor run at the same current.

Consider the case of an SST4416. It has a minimum I_{DSS} of 5 mA and a minimum g_{fs} of 4.5 mS. (The units are milli-Siemens. Older texts referred to *mhos*, ohms written backwards, and used an ohm symbol upside down. 1 Siemen ≡ 1 mho.) A bipolar transistor would have $g_m = \frac{I_C}{V_T} = \frac{0.005}{0.026} = 192\ \text{mS}$, demonstrating this 40:1 gain ratio.

JFETs are therefore advantageous for amplifiers only where the source impedance is so high that the base current loading effect or bias current noise are causing problems.

The use of JFETs as current sources is demonstrated by a family of two-terminal JFET current sources, the CR160 series by Vishay-Siliconix. These devices are optimised and characterised for use as current sources.

FIGURE 9.8C:

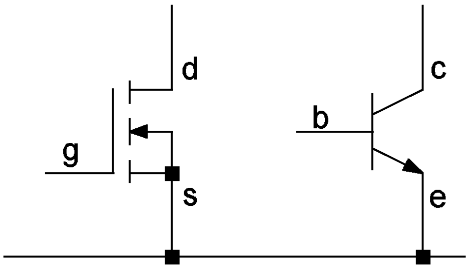


It is the upper curve that always gets emphasised in the data sheet. This is the typical impedance of the current source, with an applied voltage of 25 V. The lower curve is the more realistic set of values as a minimum impedance with 6 V applied. Note how bad the devices can be relative to their typical values. Whilst these two-terminal current sources are convenient, they do not give very good performance relative to a bipolar transistor solution. The same is evidently true of a current source that you make from JFETs in general; you will get better performance from bipolar transistors.

9.9 The MOSFET

The power MOSFET is the king of the switched-mode power supply. You may be tempted to think that the MOSFET, having an insulated gate, draws no gate current. That is the first mistake that a beginner makes. A power MOSFET can have a huge capacitive gate load to drive,[5] and when it is being switched, the current can get rather large.

FIGURE 9.9A:



Drawn like this, the NPN transistor and the N-channel enhancement MOSFET are very similar. You drive the gate [base], you ground the source [emitter] and the output is on the drain [collector]. Notice the details of the symbol. The symbol shows the gate separated from the channel, indicating that it is an *insulated gate* FET. Also the channel has been shown as a broken line, indicating that the channel does not conduct in the absence of gate drive; it is an *enhancement* device

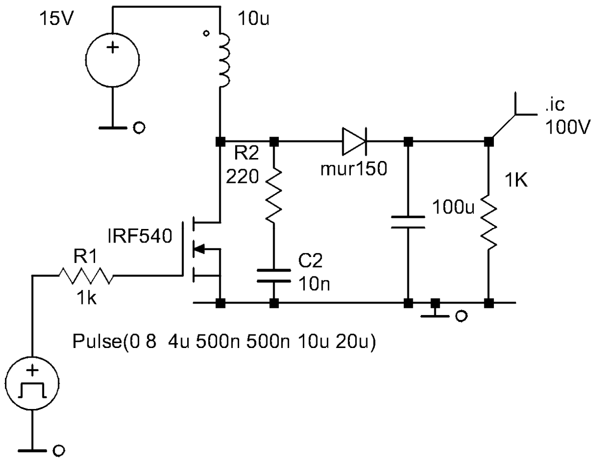
The phase of the MOSFET output is the same as that of the NPN. To turn it OFF apply 0 V to the gate (relative to the source); to turn it on apply +10 V to the gate. Different devices will start to turn on at different values, some around 2 V – 3 V, but you can apply 10 V and there is no limiting effect in the gate circuit. Indeed MOSFETs give lower ON-resistances when their gates are driven to higher voltages, the limiting value often being around 10 V to 15 V.

The datasheet parameter that tells you when the MOSFET starts to turn on is the *gate-source threshold voltage*, V_{TH}. This can have a large spread for a particular type of MOSFET. For example a BSN20 MOSFET has a range from 0.4 V to 1.8 V. The threshold voltage is also temperature dependant, so you may need to increase it by 10% for low temperature operation, or decrease it by 10% for high temperature operation. Using a simple model, if you want the MOSFET to be ON then $V_{GS} > V_{TH}$. If $V_{DS} < (V_{GS} - V_{TH})$ as well then the channel is approximately resistive, the conductance being proportional to $V_{GS} - V_{TH}$. Thus $R_D \propto \dfrac{1}{V_{GS} - V_{TH}}$ and $I_D \propto V_{DS}(V_{GS} - V_{TH})$

There are two capacitances to drive: C_{GS} and C_{DG}. As you start trying to increase the gate voltage, initially the gate load is the parallel combination of these two capacitances. However, as soon as the output voltage starts to change you get ***Miller*** feedback, increasing the effective drain-gate capacitance. This results in a characteristic slow-down of the gate voltage around the switching point. In switching devices this is bad. If the device is conducting current whilst standing-off voltage, power is being dissipated; this is inefficient. The device must not be left in this condition for very long if high efficiency is to be achieved. Driving the gate harder removes this lossy condition faster. Therefore you need a low source-impedance gate drive. Beginners think that MOSFETs don't draw gate current; now it has been established that they draw rather more than was expected!

[5] EXAMPLE: Siliconix SMP60N03 60 A MOSFET has a gate input capacitance, C_{ISS}, of 2.6 nF and a reverse transfer capacitance, C_{RSS}, of 0.75 nF.

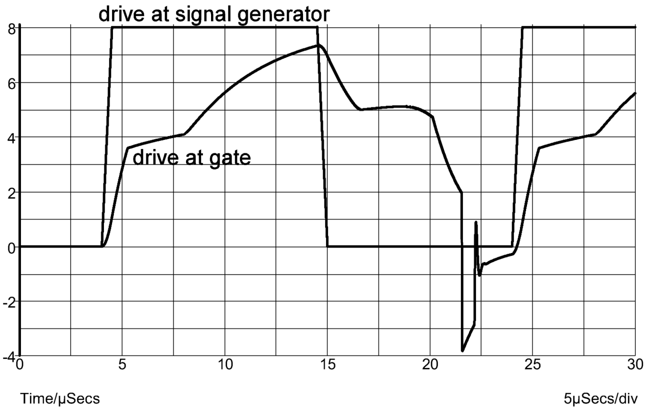
FIGURE 9.9B:



This is a simulation of a switched-mode power supply; a *boost converter*. R2 and C2 form a *snubber circuit*. The purpose is to suppress ringing on the MOSFET. This is not a power efficient technique, but is still used. R1 is the gate drive source impedance, made artificially high to demonstrate the gate drive requirements.

MOSFETs can get into parasitic oscillations just like bipolar transistors; the solution is a *gate stopper resistor*. This resistor would usually be around 47 Ω to 100 Ω, right next to the gate. Alternatively, a ferrite bead can be used to suppress the oscillations.

FIGURE 9.9C:



The drive waveform at the gate is remarkably different to the rectangular drive waveform on the other side of the resistor. The signal generator has to source and sink over 7 mA in this simulation. (Look at the instantaneous voltage difference across the resistor and divide by the 1K resistance.)

A useful application of a power MOSFET is as a high power active load to test or stress power supplies. Huge rheostats that can take tens of amps and hundreds of watts are both expensive and hard to come by. By taking a huge power MOSFET (say >500 W) and bolting it to a huge heatsink (say <0.1°C/W) you can soak up huge amounts of power from a power supply, but use only a low power potentiometer as the control device. The gate of a MOSFET is insulated, so whilst it takes a lot of current when switching quickly, the gate current for a steady gate-source voltage is very small (microamps). The potentiometer can be wired across the source and drain, with the wiper connected to the gate.

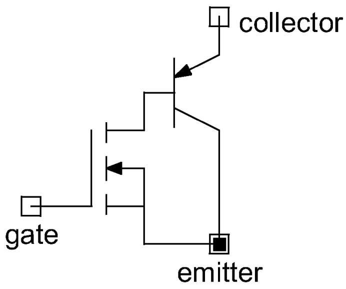
If the power supply is running at 5 V or less it may be necessary to use a bench power supply to drive the gate in order to give a gate voltage which can go all the way up to 10 V, allowing the MOSFET to turn on hard, despite lead resistances. It is wise to put a ferrite bead on the MOSFET gate lead, close to the gate, to prevent parasitic oscillation. A more general purpose solution would use an opamp sensing the voltage developed across a 4-terminal sense resistor in the source lead. This solution would give a much better defined load current. You can also buy commercial 'active loads' which look just like bench power supplies.

9.10 The IGBT (Insulated Gate Bipolar Transistor)

The *Insulated Gate Bipolar Transistor* is yet another power switching device to add to the collection. In any particular application it is not immediately apparent which type of device will give the most efficient switching operation. Although power MOSFETs do give very simple designs, bipolar transistors can give superior performance at higher voltage levels, say above a few hundred volts. The reason is that empirically, for a given die size, the ON-state resistance of a MOSFET increases faster than linearly with it its operating voltage:[6] $R_{DS(ON)} \approx R_0 \cdot V_{PK}^{1.6}$. This means that the ON-state power loss in a bipolar transistor can be considerably lower than a MOSFET. One also has to take into account the switching, bias and control losses, which is why it is not possible to state categorically that one type of device gives lower losses under all circumstances.

The IGBT is a hybrid device. It has an enhancement MOSFET as its input stage and a bipolar transistor as its output stage. Depending on the switching speed and characteristics of any particular device, it can give more efficient performance than either a bipolar transistor or a MOSFET. Check the latest available devices and see which sort of device is best suited to your application in terms of availability, cost and efficiency.

FIGURE 9.10A:



This simplified internal equivalent circuit of an IGBT shows the operation of the device quite nicely. The N-channel enhancement MOSFET makes the overall device behave as a cross between an NPN transistor and an N-channel enhancement MOSFET. The overall collector and emitter terminals seem incorrectly labelled with respect to the PNP transistor, the labels being for the overall IGBT.

The IGBT can give far superior ON-state performance compared to a MOSFET of similar die size when the operating voltage is several hundred volts or more. The only drawback is that the base-node of the internal PNP transistor is not accessible, sometimes making the turn-OFF time the largest contributor to switching losses.

Because the IGBT “output stage” is a PNP transistor, the collector-emitter saturation voltage will never be lower than the base-emitter voltage. This ‘limitation’ of 0.7 V is not a problem in reality because for switching above a few hundred volts, the ON-state voltage of MOSFETs at full current can be several tens of volts.

Another hidden loss occurs when the switching system forces a reverse voltage across the switching device. The MOSFET has an internal “body diode” which is often a poor diode, particular for a high voltage device. A more efficient external diode may therefore cut down on the switching losses. An IGBT does not have an internal body diode and can therefore utilise an external diode more efficiently.

[6] 'IGBT Characteristics', *Application Note AN983* (International Rectifier).

CH10: the opamp

10.1 The Rules

Operational amplifiers originated in analog computers and were initially based on thermionic valves {vacuum tubes}. Subsequent designs were based on hybrid transistor circuits. However, the first true monolithic opamp, the μA709, did not appear until around 1967. Now there are literally hundreds of different types available.

Operational amplifiers are easy to use from a design point of view because all the relevant specs are usually well tabulated. You should already know the basics of opamps. If you get lost, consult your elementary texts.

Now you may think that you can just design a circuit, 'plug-in' any opamp and it is done; think again! Major manufacturers produce a large number of different opamps (greater than 50). These can range under $0.30 for a cheap quad opamp to over $15.00 for an expensive single opamp. Selection of the correct opamp is therefore part of the job. Over-specify the part and you will be throwing money away. Under-specify and the circuit will drift or be unreliable to manufacture.

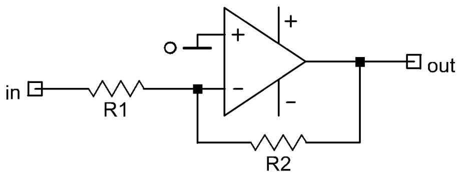
DC calculations with opamps are based on a simple set of rules. The most important rule is that the gain of the opamp is considered infinite. This means that it is not necessary to have any voltage between the inputs of the opamp in order to get a full signal swing at the output. Consider a typical opamp running on ±15 V rails and having a DC voltage gain of 1,000,000. If the output swings to say +12 V, the input signal required is only 12 μV. Thus to see how an opamp circuit works, always consider that the two inputs of the opamp are at the same voltage whenever the amplifier output is not 'hitting the power rails'; it has been said that an opamp 'likes keeping its legs together', legs in this case meaning its input terminals.

Some clarification is needed concerning 'hitting the power rails'. An opamp on ±15 V power rails will probably allow its output to swing to ±13 V, depending on the load it is driving. It is *impossible* for the output to be able to pull a load all the way up or down to the power rail. Modern 5 V CMOS opamps quote "rail to rail" operation, but this statement has to be considered carefully. There is an active device pulling the output up or down to the power rail. Any resistance in this device will give a finite volt drop on load, meaning that a heavier load {low resistance} cannot be pulled as close to the power rail as a lighter load. Read the opamp data sheets carefully to see how close to the rails your circuit load can actually be pulled.

The gain and offset effects of any opamp circuit are evaluated mathematically. The process is not at all difficult, provided you use the correct network analysis technique. You should have learned both *mesh analysis* (loop analysis) and *nodal analysis* in your earlier courses, although they may not have been named as such.

Use nodal analysis as it always produces one less variable and therefore one less simultaneous equation to solve. In the simplest cases, nodal analysis produces one equation, whereas mesh analysis produces two simultaneous equations. You need to be fluent in nodal analysis and Kirchhoff's Laws, or you will stumble over even the simplest of problems.

FIGURE 10.1A:



This simple inverting amplifier demonstrates the rules and the nodal analysis. Notice that the opamp symbol has been shown with power rails in this case. Beginners often forget that an opamp has to have power rails, since power connections are usually not shown in text book circuit diagrams. SPICE simulations also run without power rails, unless you are using opamp macro-models with power connections. Remember: *a device with power gain has to have a power source*!

The definition of the input terminals on the opamp is such that putting a positive voltage into the positive input makes the amplifier output go more positive. The positive input is therefore the *non-inverting input*. A positive input into the negative input terminal makes the output go more negative; the negative input is therefore the *inverting input*.

For the nodal analysis, remembering that the opamp "likes to keep its legs together", the negative input is roughly at 0 V because the positive input is at 0 V. The current in R1 is $\frac{V_{IN}}{\text{R1}}$, flowing towards the opamp input terminal when the input voltage is positive. The current in R2 is $\frac{V_{OUT}}{\text{R2}}$, flowing towards the opamp input terminal when the output voltage is positive. An ideal opamp takes no input current, so the currents in R1 and R2 are equal but opposite, giving:

$\frac{V_{IN}}{\text{R1}} = -\frac{V_{OUT}}{\text{R2}}$ This is rearranged to give $\frac{V_{OUT}}{V_{IN}} = -\frac{\text{R2}}{\text{R1}}$.

Having presented this simple opamp model, the next task is to add on error terms. Because the system is linear, you can use the ***superposition theorem*** and calculate each error term separately, adding all the error terms together at the end.

A real opamp needs to be given a small DC input signal in order to drive its output to 0 V; this is known as its *input offset voltage*. To model this, put a voltage source in series with the positive input of the ideal opamp. This input offset voltage will be between a few microvolts and 10 mV, depending on the type and selection grade of opamp chosen. The voltage noise of the opamp can also be conveniently included in this offset voltage generator.

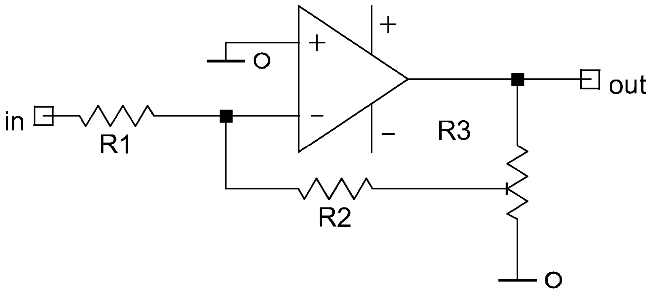
A real opamp draws *input bias current*. This could be really low for a CMOS amplifier, say a few femto-amps (1 fA=0.001 pA) or it could be as high as a few microamps for a bipolar opamp. The input bias current is the mean value of the currents going into the two inputs. Ideally these currents would be exactly equal; in practice there is an imbalance. The imbalance is known as the *input offset current*. This situation is modelled by applying a current source of $\left(I_{BIAS} + \frac{I_{OFFSET}}{2}\right)$ to one input and $\left(I_{BIAS} - \frac{I_{OFFSET}}{2}\right)$ to the other input. The bias noise current would be included in both of these bias current sources, but the noise sources would be uncorrelated with each other.

At college you will have been taught about power gain; all 'gains' are considered as power gains. When working with opamps, 'gain' is almost always used to mean voltage gain. The reason is that the input impedance of a stage is usually high (>1 kΩ) and the output impedance is usually low (<10 Ω). Power gain is therefore not of any great significance to the design process. Ideas such as 'impedance matching' and 'maximum power transfer' are therefore usually both misleading and irrelevant for intermediate opamp circuits, the exceptions being at transducer inputs and load driver outputs.

Opamps are not designed for, or suitable for, operation above about 500 MHz. Specialised fixed-gain amplifiers (usually 50 Ω input & output impedance) are then brought into play for particular frequency bands all the way up to 100 GHz.[†]

10.2 Gain

FIGURE 10.2A:



Here is a slightly more complicated inverting amplifier. The pot, R3, allows more gain without having to use too large a value for R2. This output voltage division could also be done with a fixed pair of resistors instead of the pot.

@EX 10.2.1: What is the LF input-to-output voltage gain, neglecting any opamp imperfections and neglecting the output impedance of the pot? Give your answer as a function of R1 and R2, with the position of R3 given in terms *p*. Use *p*=0 as the grounded end of the pot.

The factor by which the voltage noise of the opamp is amplified is the *noise gain*. In addition to evaluating the noise, the noise gain is also important because it is the multiplier used to work out how the input offset voltage and input offset voltage TC affect the output. The opamp voltage noise will be given on the data sheet in either $\text{nV}/\sqrt{\text{Hz}}$ or μV RMS in a certain bandwidth.

WARNING: These numbers for the voltage noise of opamps might be referred to as "noise figures" but this is inadvisable. The term ***noise figure*** is widely used by RF engineers and has a different meaning. Thus if a data sheet or text book mentions "noise figure" one has to be careful to establish, by context, whether it means "noise numbers" or the RF definition. A true *noise figure* will be in dB; values in μV and $\text{nV}/\sqrt{\text{Hz}}$ are best kept well away from the word "figure".

All of the voltage noise of the opamp can be lumped into a single voltage generator placed in series with the non-inverting input. This same voltage generator could be put in series with the inverting input instead, but it is then much easier to make a mathematical mistake.

@EX 10.2.2: Using the same rules and circuit as in the previous exercise, what is the voltage noise gain of the amplifier?

[†] eg Northrop-Grumman Velocium range of indium phosphide (InP) HEMT MMICs.

Although more gain means more noise at the output, the noise gain *referred to the input* is a more useful figure. Refer the noise to the input by dividing by the voltage gain.

@EX 10.2.3: Using the same rules and circuit as before, what is the voltage noise gain **r**eferred **t**o **i**nput (RTI)?

The voltage noise gain referred to input is given by: $\left(1+\frac{\text{R1}}{\text{R2}}\right)$. For a gain of -1 the noise gain is 2. Therefore, for a given opamp, using an inverting gain of 1 gives more than double the noise of a non-inverting gain of 1. The noise is more than doubled because on the inverting amplifier there is the *Johnson noise* of the resistors to take into account as well. As the gain goes up, the difference in noise gains between the inverting and non-inverting configurations becomes unimportant. This scheme of using an output voltage divider to boost the gain is therefore only useful if R2>R1.

Current noise in the opamp also has to be considered. Current noise is specified in terms of $pA/\sqrt{Hz}$ or pA RMS in a given bandwidth. Multiply the current noise by the source impedance to get another amount of voltage noise to add on. Thus the quietest amplifier for any given application is the one for which the combination of the voltage and current noise is the least.

These noise sources are considered as random (uncorrelated) RMS noise values. They combine by taking the Root of the Sum of the Squares of the individual RMS values; this is the RSS value. Noise *powers* add directly, and by using RMS voltages combined in an RSS fashion, you achieve the same result.

***EX 10.2.4**: Amplifier X has $4\,\text{nV}/\sqrt{\text{Hz}}$ and $1\,pA/\sqrt{Hz}$ wideband noise. Amplifier Y has $15\,\text{nV}/\sqrt{\text{Hz}}$ and $0.02\,pA/\sqrt{Hz}$ wideband noise. They are both wired as unity gain followers. For the same bandwidth:

a) which amplifier is quieter for a 1K source impedance?
b) which amplifier is quieter for a 100K source impedance?

It is important to realise that, for a given device, *the noise is always less when you run at lower source impedances*. The reason why you might need to operate with higher source impedances may be to do with the type of sensor or transducer you are using, or it may be related to the low speed of operation of a filter you are designing. The answer is to make this impedance as low as you can get it, then chose an optimum opamp.

Some reasons why you wouldn't try to minimise noise by always using resistors below a few tens of ohms are:

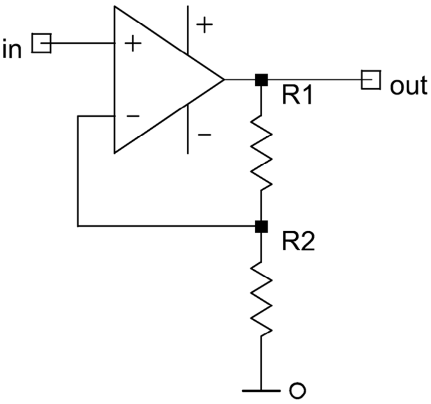
- increased power consumption
- increased non-linearity and harmonic distortion
- filter capacitors would get unreasonably large and expensive
- inability of many opamps to drive more than about ±30 mA load current

***EX 10.2.5:** Using this non-inverting amplifier as an example, decide mathematically if either or both of these two statements, both found in application notes, are correct:

a) The bias current only flows in the feedback resistor, so the resulting error is the bias current multiplied by the feedback resistor.

b) The source impedance at the inverting input is the parallel combination of the two resistors, so the resulting error is the bias current multiplied by this impedance.

FIGURE 10.2B:



Thinking of opamp circuits in terms of the *virtual earth* principle may cause trouble when trying to understand high speed amplifiers. The rules, as given in the earlier section, work well at DC, but need to be understood better as signal frequencies increase.

EX 10.2.6: An opamp having a gain-bandwidth product of 50 MHz is wired as an inverting amplifier, its non-inverting input being grounded. Its output is swinging 10 V ptp at 1 MHz (sinusoidal) due to an input stimulus to the resistor/capacitor network that surrounds it. What is the nature and magnitude of the signal on the opamp's inverting input pin?

The 'virtual earth' is not at zero volts for high speed signals, even when using a fast opamp. If there were diodes from the inverting input to ground, perhaps for overload clamping purposes, they would certainly conduct on fast edges.

If you switch feedback resistors in an opamp circuit, in order to make a switched-gain stage for example, the higher gain positions will have considerably less bandwidth. There are four possible solutions:

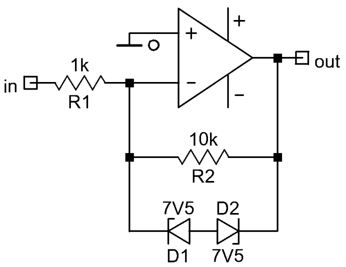
1) This stage has to have so much bandwidth, even in the lowest bandwidth gain setting, that it does not seriously limit the system bandwidth.
2) Capacitors also have to be switched in order to minimise the bandwidth change.
3) A *current feedback* opamp can be used in order to minimise the bandwidth change (this type of opamp is discussed later in this chapter).
4) The bandwidth change can be neglected as not causing a problem.

10.3 Clamping

An opamp that is driven so hard that its output 'hits' one of the power rails is severely overloaded and will usually have an unspecified recovery time. Thus you may have to characterise the specific opamp for the application in order to guarantee an acceptable recovery time. The recovery time may also vary from manufacturer to manufacturer of otherwise similar parts. For this reason clamp circuits are employed.

Another reason for using a clamp is to prevent overload or damage to a subsequent stage. A good example of this is on an ADC input. It is usually necessary to clamp the overload signals that can be fed into an ADC, but this clamp must not be allowed to affect the linearity when the signal is within the input range of the ADC.

FIGURE 10.3A:

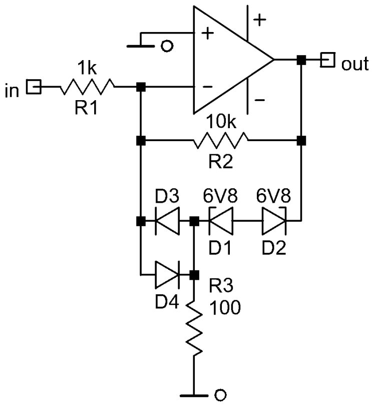


Here is a beginner's attempt at clamping an opamp-based amplifier.

***EX 10.3.1:** Using nominal values for the components and standard zener diodes (all at room temperature), roughly how high can the amplifier output rise before the gain or linearity of this circuit can no longer be guaranteed to better than 1% accuracy?

You can change the zener model used and get a different answer, but why bother? The circuit is fundamentally unsound and can be made vastly better at hardly any cost.

FIGURE 10.3B:



This is a better clamping arrangement. The current in the zener diodes is defined by R3 when the clamp starts to operate; the leakage current across R2 is very well defined. Note that the amplifier has to be *unity gain stable* for this clamp arrangement to work correctly. When the clamp is active, the small-signal resistance from output to input may only be a few tens of ohms, giving almost 100% feedback.

At high speeds the zeners will take time to charge using this scheme, allowing excessive overshoot. A more complicated scheme keeps both zeners biassed all the time, then uses low-leakage reverse biassed diodes to define the clamping points. This gives less overshoot on fast overload. However, the design time and parts count is increasing. A better alternative is to use a *clamp-amp* if you only need ±5 V operation.

A clamp-amp is a modern opamp variant with externally set upper and lower output thresholds. These are available from several manufacturers and are the parts of choice for driving ADC inputs. They limit the overdrive into the ADC inputs and yet have a sharply defined 'knee', minimising the non-linearity near the clamping levels. Read the specs of the latest devices to see if they are suitable for your application. The 'softness' of the knee is the key performance point that you need to watch out for.

Another problem that you may come across is damage to the input stage of the opamp. When thinking of the opamp as a building block, it is easy to forget that there are transistors inside. Ordinarily opamps are short-circuit protected at the outputs and it is easy to think of them as relatively indestructible. Input stages are another matter, however.

Take a simple differential pair amplifier made in bipolar transistors, for example. If you applied a 10 V signal between the inputs you would expect one of the transistors to get damaged; the one which got more than 6 V reverse bias on its base emitter junction would certainly be 'unhappy'. The result could be increased noise and/or increased bias current.

The moral of the story is that the maximum differential voltage limit of the particular opamp you are using must be checked. It may be a JFET opamp with a ±30 V differential voltage limit and therefore no problem. It may be a ±15 V differential limit; it may instead be a low-noise bipolar input that can only take ±0.7 V differential input voltage because of internal clamp diodes. Read the data sheet and see if you have to take extra precautions.

When using dual or quad opamps, it is important to keep all sections of the package running in their linear range. If you are only using three opamps in a quad package, for example, wire the last opamp as a ×1 buffer connected to 0 V (provided it is unity-gain stable). This rule applies to digital gates as well, particularly for CMOS devices. Always define what the input should be to prevent oscillations and unexpected behaviour.

10.4 Instability

Data sheets sometimes contain clauses such as ‘for use at gains greater than 5’ or ‘minimum stable gain 5’. Another way of writing this is to say that the amplifier is not *unity-gain stable*. This last statement is never found in data sheets, presumably because the manufacturers do not wish to stress the fact that the opamp will oscillate if wired in a unity gain configuration. After all, who would want to use an “unstable” amplifier in their design? Another word that is found in data sheets is *decompensated*. In this case the *de-* prefix meaning to remove or take away (the compensation). These are all ways of saying exactly the same thing, without ‘worrying’ the novice.

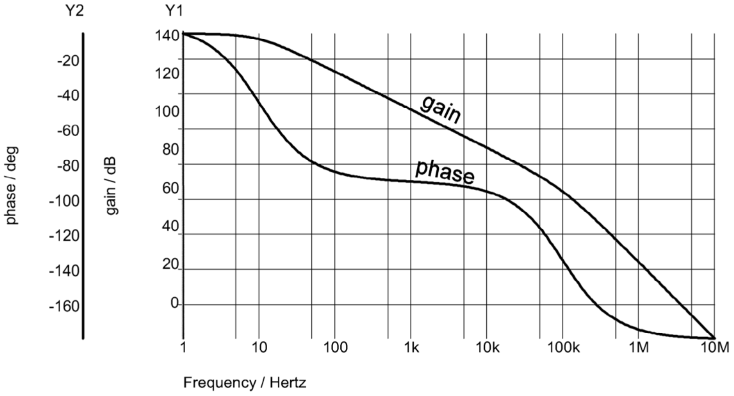
At this point you may have to get out your control theory books and do some revision. Unless you understand these basics, the techniques that follow will seem a bit strange. The fact is that if you need a gain of ×10 in an amplifier stage, you will always get better performance if you use the de-compensated (unstable) version of any particular opamp family. Lots of gain-bandwidth product is necessary in the amplifier in order to produce lots of bandwidth in the finished stage.

***EX 10.4.1:** Why is a ×1 non-inverting opamp configuration more likely to oscillate than a ×1 inverting opamp configuration?

If you think of an opamp in terms of a macro-model with a certain DC gain and a single-***pole*** roll-off, you will go astray. Opamps do not behave in this way and simulations based on such a simple macro-model often give misleadingly good results. You should usually consider an opamp in terms of a two-pole macro-model. The *second*-pole is the problem and is often hidden from direct observation.

You will know from your control theory ***Bode*** plots, that if the slope of the ***loop-gain*** crosses the 0 dB point at a rate of 40 dB/decade, that system is unstable.

FIGURE 10.4A:



In this simple model, the first pole is at 10 Hz and the second pole is at 100 kHz. There is a good ***gain margin***, but virtually no ***phase margin*** for operation with a closed-loop gain of unity (by using a feedback gain of 1).

Remember that a pole introduces a 45° phase shift at its corner frequency. Thus if you position the second pole to give 0 dB overall gain at the pole, you will get 135° phase shift at this frequency, resulting in a 45° phase margin at unity gain. In order to get 45° phase margin with the amplifier graphed above, unity loop-gain would have to be at 100 kHz. Thus you would need to give it a gain of more than 700.

You can see from the figure why it is necessary to quote the gain-bandwidth product at a particular frequency. The gain-bandwidth product assumes a single-pole model for the opamp. When the second pole is visible, as in the above figure, the unity gain cross-over frequency is not equal to the gain-bandwidth product; it is quite noticeably lower. In this example the gain-bandwidth product is 100,000,000 between roughly 10 Hz and 100 kHz. However the gain becomes one at 3.2 MHz and not 100 MHz. Yet again the second-pole is the one to watch out for.

Any capacitive loading on the output of the opamp will form yet another pole with the output resistance of the amplifier stage. The additional phase shift will always reduce the phase margin. You will see figures in opamp datasheets indicating maximum load capacitance, but approaching this maximum value will cause the pulse response to overshoot and ring excessively. Thus you should be careful about allowing too much capacitive load directly on the output of an opamp; more than 100 pF can be too much. Having said that, there are newer opamp designs which are specifically designed to drive several nanofarads of load capacitance. Also one can drive capacitive loads using a more complicated feedback arrangement of resistors and capacitors.

In switched-gain amplifiers, it is quite usual to change the feedback resistor to switch the gain. In an inverting configuration you might have a 10K input resistor and decade switched feedback resistors in the range 1K to 1M giving inverting gains of ×0.1, ×1, ×10 and ×100. What happens is that the bandwidth drops off at the ×10 and ×100 gains because of the finite gain-bandwidth product of the amplifier. On the other hand, the amplifier is much more likely to oscillate at the gain of ×0.1 because the feedback amount is so high.

In this application, the amplifier has to be unity gain stable unless you play games with the opamp frequency response on the different ranges. One way of doing this is to also switch the *external compensation capacitor*. Not all opamps have these, but the idea is that the capacitor is used to adjust the gain-bandwidth product of the amplifier. Since you want less gain-bandwidth product on the low gain ranges, it is convenient to switch-in more compensation capacitance on the low gain ranges in order to optimise both the low gain and high gain positions.

Although a ×100 amplifier is more stable than the ×0.1 amplifier in terms of the simple opamp model, *unintentional* feedback also occurs due to power supplies and control line pickup. A high gain and/or high bandwidth amplifier is *much* more susceptible to this type of unintentional feedback. If you make a single-ended 1 GHz amplifier with a voltage gain of ×100 it is almost guaranteed to oscillate. It is incredibly difficult to stop the output signal coupling back to the input. There will be one or more points where the loop phase shift is exactly 0° and the loop-gain is greater than unity; an oscillation will develop at this frequency.

Solutions include:

- Using differential input and/or differential output stages to balance up the signals.
- Individual local decoupling of the power rails to each amplifier in a chain of amplifier stages.
- Not running control lines near the output or input tracks
- Decoupling control lines so that they cannot carry modulated signals.
- Putting shielding cans over the amplifier so that the stray coupling from input to output is minimised.

Layout effects are not shown on a circuit diagram, will not show up in a simulation, and will not be resolved without physical manipulation of the PCB. Sometimes a pulse response problem can be fixed by an extra decoupling capacitor on a long control line. Sometimes the track needs to be re-routed. The correct solution is to minimise the possible interaction before you get a problem, routing such tracks around the outside edges of the amplifier and not through the middle of the amplifier.

It is quite usual to have opamps running on ±12 V or ±15 V rails relative to some 0 V signal ground. The reason for this is that often neither the inputs nor the outputs are allowed to get very close to the power rails. Specs such as bias current and common-mode rejection ratio are specified over a limited range, the ***common-mode input range***, which may include only one of the power rails or it could exclude both power rails. Sometimes the common-mode input range is specified as being from $V^- + 3$ (3 V above the negative rail) to $V^+ - 2$ (2 V below the positive rail). An amplifier of this type would be no good on a single 5 V power supply because the common-mode input range would be violated regardless of the input voltage!

Over the years 1995-2000 many new opamps became available, characterised specifically for use at 5 V, 3 V and even lower power rails. They were specified and advertised as having 'rail-to-rail inputs'. Nowadays you can even find amplifiers where the common-mode input range exceeds the power rails! Check the opamp datasheet to make sure that the common-mode input range is suitable for your application.

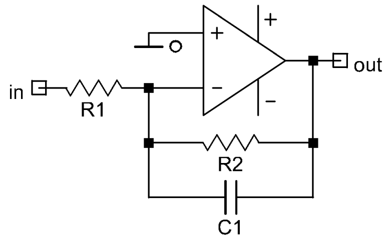
An opamp does not 'know' what absolute power rails it is running on, it only knows about difference in voltage. Thus a ±5 V specified device can be run from +95 V to +105 V rails and it will be unaffected (provided that its inputs and output are also up in this range).

Allowing an opamp's inputs to go outside of the defined common-mode input range can cause one of two things to happen: the input bias current and/or input offset voltage may get dramatically worse, or the amplifier may *phase invert*. Having input bias current and input offset voltage getting dramatically large is not normally a problem when the amplifier is being run outside of its normal operating range. Phase inversion of the output, however, is always very unpleasant.

Phase inversion is where an increasing input voltage, which was causing an increasing output voltage, suddenly causes the output voltage to start decreasing! A heavy overload can then re-appear in the linear range of the next device in the signal path. It is essential to prevent such occurrences for any sort of instrumentation application. The remedy is either input clamp circuitry or a replacement amplifier type.

10.5 Basic Amplifiers

FIGURE 10.5A:



This is a basic inverting amplifier. It is the easiest configuration in which to do *offset nulling* because there is no common-mode signal to deal with.

EX 10.5.1: You are required to completely null the ±5 mV offset voltage of this amplifier using a pot. (Neglect bias current errors.) Design the circuit and give the values necessary. Power rails of ±12 V (±5%) are available. Use realistic component tolerances.

Do not use the input offset nulling points of an opamp to null offsets generated in the rest of the circuitry. When you null the input offset of an amplifier you may improve its TC as well; you are removing an imbalance. If you null-out errors in other parts of the circuit you will be *creating* an imbalance which then causes a worse TC.

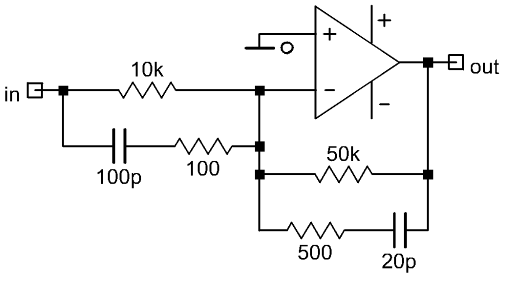
C1 in the figure above is needed to counteract the pole created by R2 driving the input capacitance of the amplifier. C1 will be somewhere in the region of 1 pF -10 pF. It is the Thévenin equivalent R1//R2 which drives the input capacitance. If R1//R2 is around 1 kΩ it may be possible to omit C1. However if R1//R2>100 kΩ then C1 will almost certainly be needed, depending on the gain-bandwidth product of the opamp being used.

EX 10.5.2: How much phase shift is caused by R1//R2=100 kΩ driving a 4 pF amplifier input capacitance at 1 MHz? Consider the amplifier input capacitance to be a capacitor to signal ground.

That realistic example illustrates the problem with the input pole of an inverting amplifier. Having included C1, the bandwidth is now reduced. If this bandwidth reduction is unacceptable, a capacitor has to be put across R1 to make the overall bandwidth higher. The medium frequency gain is then set by the capacitors and not the resistors.

For performance >20 MHz the capacitors will need small resistors in series with them to stop the capacitors becoming too much like short circuits. These small resistors might be in the region of tens to hundreds of ohms.

FIGURE 10.5B:



This circuit shows the complete wideband circuit. To obtain the flattest possible frequency response, the low value resistors will not necessarily have the same ratio as the high value resistors. The low values often need to be adjusted to allow for finite gain in the amplifier and the finite output impedance of the previous stage.

Manufacturer's data sheets for high spec opamps can give harmonic distortion figures that extend up to incredible levels such as –110 dBc. You might reasonably ask how such performance can be measured. Well a pure sinusoidal source is a good starting point, and this can usually be improved by putting a narrow band filter in series to filter out some of the harmonic content. However, one trick is to make a unity gain inverting amplifier using the opamp. Now the incoming signal and the amplifier output signal can be resistively subtracted, causing both the incoming fundamental and its harmonics to be nulled out. The remaining signal will have a much higher proportion of the amplifier's harmonic distortion than the output signal on its own would have.

Suppose you apply a 1 V signal and you get approximately 1 V out of the amplifier as well. These can probably be subtracted from each to leave a signal of not more than 10 mV. If you adjust the components in the resistive subtracting network carefully, maybe you can null the signal down to 1 mV, but you may need to use some sort of cable delay or other phase shifting network to get the best possible null. Having nulled the fundamental by 40 dB or more, you have improved the resolution of the harmonic measurement by this same amount. Thus an ***FFT*** analyser that itself had –60 dBc range would now be measuring distortion in the amplifier at the level of –100 dBc.

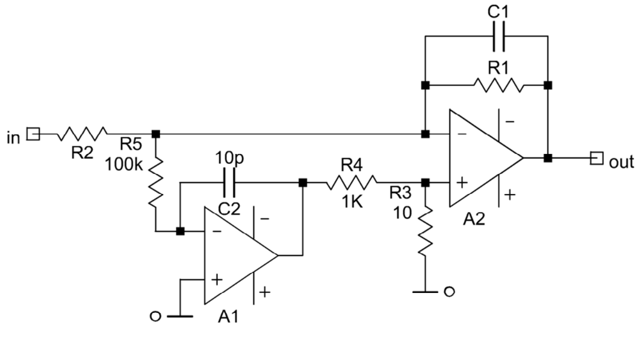
Two-tone ***intermodulation distortion*** testing is also a useful way of testing for non-linearities. It has one great advantage compared to direct harmonic distortion testing; there is no need to produce a harmonically pure test signal in the first instance.

10.6 Compound Amplifiers

It is quite normal that existing opamps are not good enough for a demanding application, despite the fact that there are hundreds to choose from. One is optimised for high bandwidth, but has poor offset characteristics. Another has excellent low noise, and low offset, but has hardly any bandwidth. Or perhaps you want a power opamp, and none have sufficiently low offset-voltage temperature coefficients. The solution is to make a compound amplifier using two opamps.

Consider offset-nulling a fast amplifier. It is possible to drive an offset-nulling signal into the offset-null pins of the amplifier. The only problems are that this may cause a worsened power supply rejection ratio (***PSRR***) or the amplifier may not have offset-nulling pins. These problems are overcome by feeding the correction signal into one of the amplifier's inputs.

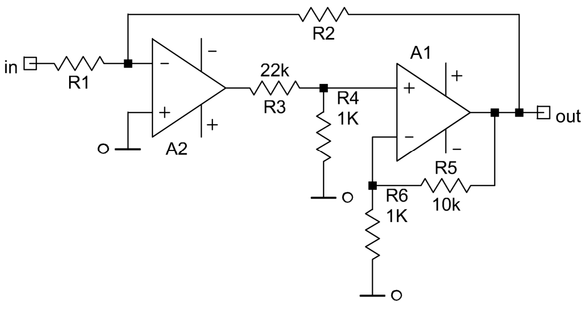
FIGURE 10.6A:



Notice that the design complexity has increased considerably. The components R5-C2 prevent low speed A2 from experiencing HF signals beyond its capability. Attenuator R3-R4 prevents A1 from being 'blown up' if the compound amplifier becomes overloaded. An alternative is to put a pair of inverse-parallel diodes in place of R3.

Another 'traditional' compound pair uses a power stage to drive a heavy load with a small-signal opamp to provide the gain, linearity and noise performance.

FIGURE 10.6B:



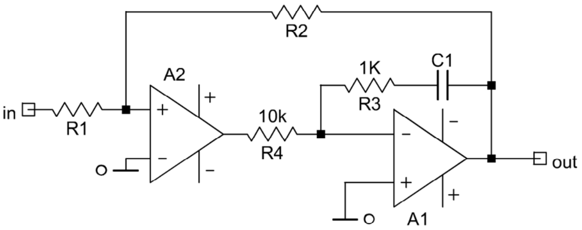
A2 is the precision opamp and A1 is the power output stage. A1 could also be a unity gain transistor output stage and the situation would be similar. I have arbitrarily given the output stage a gain of 11. This could be anywhere from 1 to 100. Notice that I have attenuated the output of A2 by 1/23, this attenuation needing to be greater than the gain defined for A1.

If A2 is unity gain stable and you add more gain, in the form of A1, then the combination will be unstable if the overall gain of the compound circuit is low. If you reduce the gain by the same amount as that introduced by A1 then you will still have lost gain/phase margin; A1 will definitely introduce additional phase shift. The exact amount of attenuation introduced depends on the overall desired gain. If the overall gain is 100, as set by R2/R1, then omit the attenuator R3-R4 completely. In fact in this case it would be advantageous to also increase the gain of A1 to between 20 and 50. On the other hand, if you want to switch the gain of the compound amplifier by switching R2, consider switching R4 to maximise the available loop-gain.

It may seem tempting to use a decompensated amplifier for A1, since it is apparently running at a gain of 11. In feedback terms, and hence in terms of the stability, A1 is actually running at a gain of 10; you may see this referred to as running at a noise gain of 10. The problem with using a decompensated opamp is that there may be insufficient gain/phase margin within the overall outer loop if the gain required is low. It is therefore inadvisable to use decompensated opamps within a compound amplifier like this.

If the system requires gain of more than 100, it may be preferable to use a compound amplifier as shown above, rather than use two separate stages. Two stages requires two pairs of gain setting resistors; the result will therefore be less accurate than using a single compound-stage having only two gain defining resistors. One drawback to this idea is that the gain may be so high that any slight feedback may cause oscillation, or at the very least peaking.

FIGURE 10.6C:



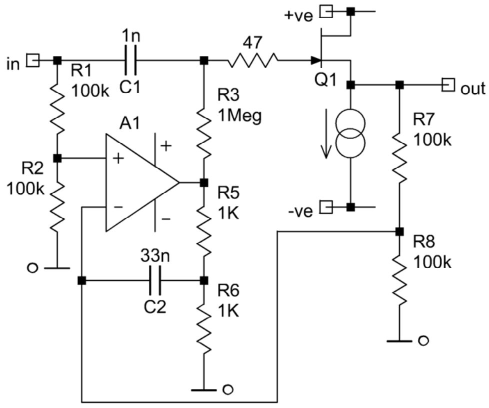
Here the capacitor in series with R3 boosts the loop-gain at lower frequencies. Make sure the gain of the A1 stage is reduced to 11 well before the unity gain crossover frequency of the compound amplifier.

High loop-gain is essential to maintain accuracy and to minimise harmonic distortion. If you need a gain of 100 from a single stage then you ideally want an amplifier that is only stable at a gain of 100. An amplifier that is stable at a gain of 1 will have much less loop-gain available for harmonic distortion reduction. The compound amplifier works well in such applications.

A compound amplifier will also be useful if the load resistance is low, say below 100 Ω. In this case there will be thermal feedback from the output stage of the opamp to its own input. This thermal feedback effect is minimised by using a compound amplifier as above.

Those were very straightforward compound designs because the compound amplifier was inside the overall feedback loop. The high open-loop gain meant that no difficult flatness problems were encountered. When one path is AC coupled then the problems really start.

FIGURE 10.6D:



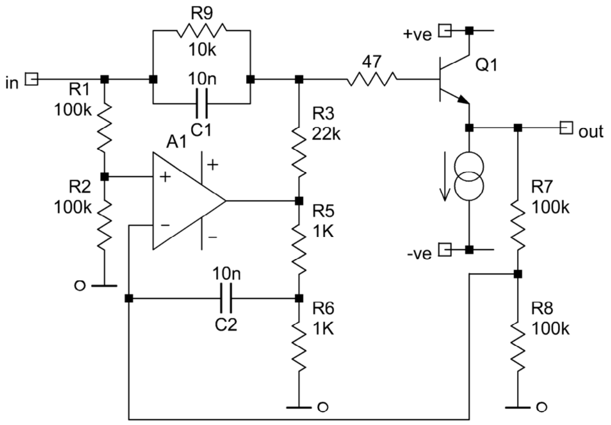
This is a high speed JFET buffer, stabilised by an opamp to take out the offset, drift and LF noise problems associated with the JFET. There are three precision dividers: R1/R2, R5/R6 and R7/R8.

The problem being solved here is how to get a high impedance buffer at hundreds of megahertz which still has excellent performance at DC. The use of a MOSFET instead of a JFET for Q1 can extend the performance up to a gigahertz.

This circuit never gives a truly flat response and the matching of the resistor ratios is essential to getting any reasonable sort of performance. Don’t expect to be able to adjust this sort of circuit to a flatness of better than ±0.1% from DC to 1 MHz. Despite what the simulations say, the stray time-constants and non-constant input impedance of the JFET will conspire to make the response less than ideal. It is also impossible for the JFET to actually give a voltage gain of 1, due to the load. If the JFET gain is down say 2%, this needs to be compensated for by reducing the value of R7 by 4%.

A particular problem with this scheme is the signal seen by A1. It is being subjected to full bandwidth signals on both inputs. A1 will inevitably run into slew-rate limit problems which may cause a frequency dependant DC offset. The signal into the A1 inputs really needs to be rolled-off before it causes trouble. This reduces the input impedance of the buffer and further reduces the flatness.

FIGURE 10.6E:

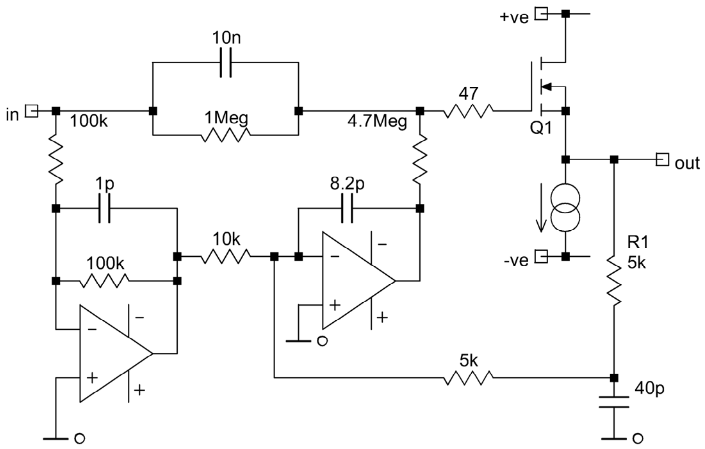


You can make the same sort of buffer using a bipolar transistor. In this case the resistor driving the base of the transistor needs to be somewhat smaller, otherwise the base current noise will be too great. The scheme is significantly improved by using R9. By reducing the need for the opamp correction, the flatness errors are more than halved. R3 would ideally be several times greater than R9. The limitation being the power supply rails of

A1. You have to supply 0.7 V of offset on R9 in order to compensate for the base-emitter drop on Q1. This limits R3 to not more than about 10×R9.

This dual path concept works with any amplifying device: JFET, MOSFET, bipolar, GaAs FET, HJBT &c.

FIGURE 10.6F:



Using inverting amplifiers, the problem of slew rate limiting in the first opamp is solved. The cost is an extra opamp, but this is balanced against the reduced need for matching in the resistors. Again R1 will need to be reduced by a few percent to compensate for the gain loss through Q1.

EX 10.6.1: You are building a new two stage amplifier and you have two previously designed amplifier modules, one with 5 MHz bandwidth and one with 50 MHz bandwidth. Both are equally quiet in terms of $nV/\sqrt{Hz}$ and $pA/\sqrt{Hz}$. They have equal input impedances and equal LF gains. Engineer *A* says you should put the low bandwidth amplifier first in order to bandlimit the noise from the signal source, thereby minimising the noise. He says that if you put the higher bandwidth amplifier first you are just amplifying the noise. Engineer *B* says they should be the other way around, but is incoherent as to why. Engineer *C* says the optimum noise performance will be achieved by having equal bandwidth in both stages and that a new amplifier module design should be undertaken. Which engineer should you agree with or does it not make any significant difference?

10.7 Differential Amplifiers

Differential amplifiers are very important for removing noise from a measurement. More specifically they are important for removing ***common-mode*** noise. In any control or instrumentation system you ideally want to amplify the signal and attenuate the noise in order to give the best possible representation of the signal. You can filter the signal if there is noise of higher or lower frequencies compared to the signal, but another thing you can do is to get the highest possible amount of the signal to start off with.

In order to get an electrical signal you always need *two* connections. Imagine walking up to the system under test with two leads attached to a moving coil meter. If there were no AC signals in the world, and there was no leakage from your meter leads, this would be a truly “floating” measurement. In other words, the reading on the meter would be correct regardless of the potential of the measured system with respect to … the meter.

If current flows down the meter leads to anywhere other than the meter, the measurement will be incorrect. At DC this would be due to leakage resistance, but for all alternating frequencies current could flow by capacitive coupling, magnetic coupling, or radiative loss. Thus making a measurement, even with a notionally floating device, is not necessarily trivial.

Suppose you are measuring the power rail on an opamp. You expect it to be +15 V. You clip one lead of the DVM onto the 0 V power connector to the board. There could easily be a few tens of millivolts, or even a few hundreds of millivolts, between this 0 V and the 0 V near the opamp. Even if you are only checking for +15 V ± 500 mV, the result could be wrong if there are several amps flowing in the power lines. The 0 V connection to the meter should either be made near to the opamp, or an additional measurement should be taken of the difference in voltage between the incoming 0 V and the 0 V near to the opamp.

If you want to look at the noise on this opamp power supply you have to be much more careful. You may now be looking for noise below the 10 mV level; the 0 V reference point will now be critical. Again the reference 0 V point ideally needs to be within a few centimetres of the opamp.

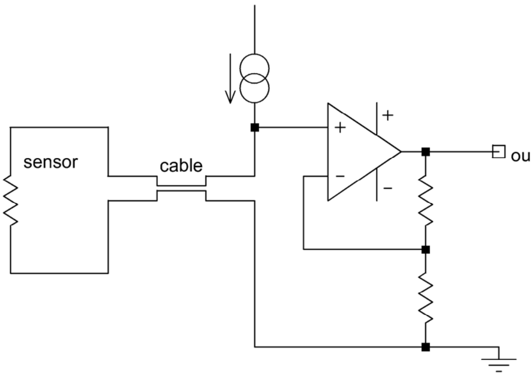
Consider the problem of monitoring a sensor in some remote location, possibly tens of metres away. Inevitably the leads will pick up all sorts of electromagnetic interference on their way back to the measurement device. Suppose it is a two-wire sensor; the two obvious things to do are to put both wires in a screened sheath or to use a twisted pair.

***EX 10.7.1:** Hint, think in terms of the physics of this problem.

a) What is the benefit of using a twisted pair?
b) What is the benefit of using a screened sheath?

Having used a screened twisted pair to get the best possible shielding, the noise problem remains. This is the system:

FIGURE 10.7A:



I haven't explicitly drawn in the interfering sources on this diagram. They will never be drawn on any real world diagram that you see either! If there is equal capacitance to both sides of the cable from some sort of interfering AC source, one side gets shunted to ground and the other side appears at the amplifier input as noise.

This noise is *avoidable*, as it originated from a common-mode source. You can therefore modify the circuit to reject {ignore as far as possible} the common-mode signal. You need to use a balanced input stage; a *differential amplifier*. A differential amplifier is defined as an amplifier with two signal inputs and one signal output, where the difference between the two inputs is amplified, and the mean value of the two inputs is ignored (rejected).

If the two signal inputs are V_1 and V_2 then the difference, the *differential-mode* *signal*, coming into the amplifier is $V_D \equiv V_1 - V_2$, on the assumption that V_1 is applied to the non-inverting input. Notice that the difference voltage is evaluated at each new moment in time. The difference voltage is not an RMS voltage obtained by the difference between two RMS voltages. The common-mode signal is $V_C \equiv \frac{V_1 + V_2}{2}$, and again it is evaluated continuously. The output signal is some linear combination of these common-mode and differential-mode voltages.

Specifically, $V_O = G_D \times V_D + G_C \times V_C$. Where G_D is the *differential-mode voltage gain* and G_C is the *common-mode voltage gain*. By definition, the differential amplifier is supposed to be amplifying the differential input and ignoring {rejecting} the common-mode input. Thus ideally G_C should be zero.

The formula can be re-arranged to more clearly express the measurement problem:

$$V_O = G_D \left(V_D + \frac{V_C}{G_D / G_C} \right) = G_D \times V_D \left(1 + \frac{1}{CMRR} \times \frac{V_C}{V_D} \right) \qquad \text{dB are not used here}$$

The common-mode rejection ratio (CMRR) would ideally be infinite, but in practice it is the spec you are trying to improve in any particular differential amplifier. CMRR is usually expressed in dB form, where every 20 dB improvement means 10× less effect due to the common-mode signal. In the formula given above, the CMRR is in ratio form, not dB form.

A standard voltage-feedback opamp is a differential amplifier. The trouble is that its differential-mode gain is too high (eg ×1,000,000) and uncertain (eg ±10 dB) to use without feedback. You might want a gain of 100× with an accuracy of 0.1% for example.

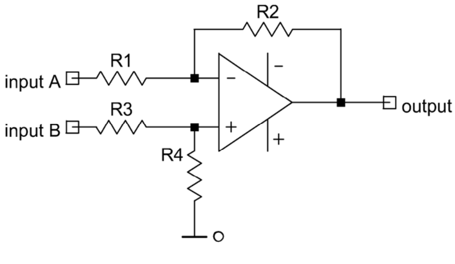
Here is the first contender for the position of a differential amplifier with a known, *stable* gain; stable, in this case, meaning not drifting with time or temperature, and no oscillation either.

***EX 10.7.2:** R1 and R3 are nominally equal to each other. R2 and R4 are also nominally equal to each other. Neglecting the deficiencies in the opamp and using the symbols:

$$\frac{R_2}{R_1} = G \text{ and } \frac{R_4}{R_3} = G(1+\delta):$$

a) What is the common-mode gain?
b) What is the differential-mode gain?
c) What is the common-mode rejection ratio?

FIGURE 10.7B:



Note that the ratio of R1/R2 could have been matched to the ratio R3/R4, without having R1=R3. This gives a non-symmetric loading at the inputs and is not recommended.

The exercise demonstrates a very important point. It is far easier to achieve high CMRR when there is gain in the overall amplifier. This fact can alternatively be expressed by saying that the requirement for matching in the resistors is reduced, for a given CMRR, when there is gain in the overall amplifier.

$$\boxed{CMRR_{dB} = 20 \times \log_{10}\left(\frac{1+G}{\delta}\right) \text{ dB}}$$

At a gain of 1, a CMRR of 80 dB requires δ to be 0.02% (200 ppm). This value of δ could be achieved using two ratio matched pairs of 0.01% or 4 individual 0.005% (50 ppm) resistors. This is quite a difficult spec to maintain. A pot could be used to trim one of the ratios to match the other ratio, but maintaining this ratio with both time and temperature is more expensive.

The value of δ required for 80 dB CMRR at a gain of ×100 is only 1.01%; this could be achieved using 0.25% individual resistors, an easy requirement to meet. Thus it is much easier to attain high CMRR on small signals. On large signals there will not be enough power supply range available in the opamp to give large amplification.

The previous circuit is a widely used industry standard configuration. Unfortunately the input resistance is low if you use low valued resistors, say <10 kΩ. But using high valued resistors, say >100 kΩ, gives excessive Johnson noise in the resistors, increased bias current noise from the opamp, and big problems with stray capacitance. These problems are solved by using a buffer stage first.

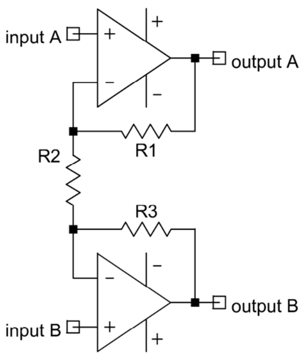
Such a buffer stage is a two-port network with two signal inputs and two signal outputs. In other words it is a differential-input, differential-output buffer. Such a buffer is better known as a *balanced* or *fully differential* amplifier. What would really be nice would be a passive device which attenuated the common-mode signal, whilst leaving the differential-mode signal alone. Unfortunately, no such device exists at DC, although a transformer performs this function over a limited band of AC frequencies.

For the fully differential amplifier, $CMRR = \frac{G_D}{G_C}$, with this CMRR being evaluated by driving into an ideal differential-to-single-ended amplifier. The instantaneous common-mode output is half the sum of the individual instantaneous output voltages. The instantaneous differential-mode output is the difference between the individual instantaneous outputs.

***EX 10.7.3:** For this scheme to work optimally, R1 and R3 are nominally equal. Assuming ideal opamps, and using the notation; $\frac{R_1}{R_2} = G$, $R_3 = R_1(1+\delta)$, calculate the following values for low frequency signals:

a) The common-mode gain.
b) The differential-mode gain.
c) The common-mode rejection ratio.

FIGURE 10.7C:



This is a perfectly balanced stage, where both amplifiers see the same load as each other at all frequencies. It is best done with a monolithic-dual opamp, since very close matching between the two amplifiers is then achieved. This circuit improves the signal-to-noise ratio by amplifying the signal, but not amplifying the common-mode noise.

There is no critical matching required between the resistors.

Increasing the differential-mode gain again improves the common-mode rejection ratio. The combination of this input stage and the previous differential amplifier gives the industry standard *three opamp instrumentation amplifier*.

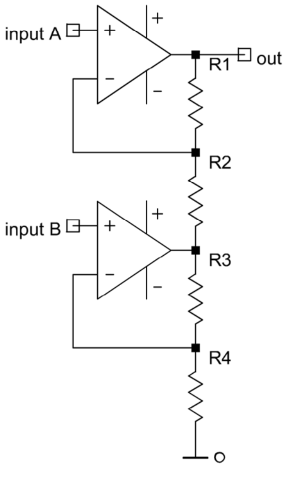
With a fully differential amplifier, like the above buffer, the outputs are a linear combination of the common-mode signal and the differential-mode signal. The maximum limit of this combined signal is an additional input constraint.

Transistorised fully differential amplifier stages can be cascaded to give CMRR values well in excess of 140 dB at low gain without the need for critical resistor matching. The CMRR values do not increase indefinitely, however, because there are actually four gain terms to consider. A common-mode input produces a common-mode output and a differential-mode output. Likewise for the differential-mode input. These additional "cross-terms" are made small by tight matching, but cannot be neglected when cascading stages since they ultimately limit the maximum achievable CMRR. The mathematics for this is explained fully in a specialist treatise.[1] Unfortunately this type of cascaded transistor stage does not have the excellent linearity and harmonic distortion qualities that are often required of modern amplifiers, so the use is highly specialised.

Just don't make your own instrumentation amplifier if an 'off-the-shelf' instrumentation amplifier will do the job. You can buy an instrumentation amplifier with a minimum CMRR of 100 dB at a gain of 10 and a bandwidth around 700 kHz as a standard item at low cost. You would be wasting time, money and PCB space by doing your own design. First check semiconductor manufacturers standard parts to see if they make a monolithic solution to your instrumentation amplifier requirement. If not, then go ahead and design your own.

There is a popularised simplification of the standard three opamp instrumentation amplifier circuit known as the *two opamp instrumentation amplifier*.

FIGURE 10.7D:



@EX 10.7.4: This circuit saves one opamp and three precision resistors compared to the three opamp instrumentation amplifier.

a) What are the requirements amongst R1, R2, R3 and R4 to make an instrumentation amplifier?

b) Why is the three opamp version still used? Hint: think about matching, phase shift and the frequency response of the amplifiers.

Although you have worked out how to achieve high CMRR by matching resistor values, I would not want to leave you with the false impression that simply by matching resistor values, and capacitor values for higher frequency circuits, arbitrarily high CMRR can be achieved. The two key facets of the design which will ultimately limit the CMRR are

[1] R.D. Middlebrook, *Differential Amplifiers* (John Wiley and sons, Inc., 1963).

harmonic distortion and phase shifts. Neither of these problems can be cured by adjustment of pots wired into the circuit to adjust the main resistor values. One then has to resort to clever compensation networks and more careful circuit layout techniques.

Furthermore, when high impedance attenuators are involved, stray capacitances mean that extra phase shifts appear in the circuit so that low frequency nulling cannot be accomplished just by adjustment of the resistors, and high frequency nulling cannot be accomplished just by adjustment of the capacitors. Hence more complicated nulling schemes may be needed that effectively change the resistive values at different frequencies.

***EX 10.7.5:** An ideal differential amplifier is connected to a signal source via test leads of unequal lengths, the difference in the wire length being 20 cm. What is the resulting CMRR at:

a) 100 kHz ?
b) 10 MHz ?

@EX 10.7.6: An ideal differential amplifier is connected to a source via equal length leads, but in one lead a single-pole low-pass filter has been inserted. The single pole filter has a 3 dB bandwidth B (Hz). What is the resulting CMRR at:

a) $B/100$?
b) $B/1000$?

It is possible to compensate for a phase shift by adding an equal phase shift to the other input. It is also possible to compensate for a specific harmonic by adding-in a suitably scaled and phase shifted signal, but this is likely to require adjustments for both the amplitude and phase. The best solution to the distortion problem is to reduce the amplifier load in order to minimise the distortion, to use better amplifiers which distort less, or to use an additional output stage which is included within the feedback loop of the final stage.

Even though your amplifier may only have a bandwidth of a few kilohertz, don't therefore assume that it will be immune to strong RF fields from mobile telephones and other similar sources. A poorly designed pressure transducer or temperature sensor amplifier, for example, can even give full scale deviations when approached by a mobile phone. The type of opamp used is critical. If there are protection diodes used internally across the opamp inputs, these can act as RF detectors and produce a DC signal which then gets amplified by the low frequency gain of the circuit.

@EX 10.7.7: Because of the high signal levels involved, it has been necessary to divide down both inputs to a differential amplifier before doing the differential-to-single-ended conversion. Assuming a perfect back-end differential-to-single-ended converter, what is the resulting CMRR due solely to the mis-match in attenuation between the two attenuators. Represent one attenuator as a gain of G and the other as $G(1–\delta)$, where $G < 1$.

10.8 Current Feedback

With ordinary [voltage feedback] opamps, engineers have become used to estimating the closed-loop bandwidth by taking the gain-bandwidth product of the amplifier and dividing by the desired closed-loop gain. If the amplifier has a GBW product of 10 MHz then 100 kHz bandwidth is estimated for a gain of 100. [Remember that 'gain' in opamp circuits always means *voltage gain*, unless otherwise specified.]

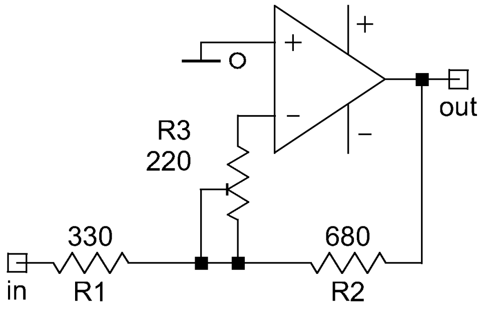
Current feedback amplifiers break those rules and have caused some engineers to have problems. Current feedback amplifiers are 'new' in the sense that they have been pushed onto the commercial market heavily since around 1988. The problem, as with any new device, is that the tricks that have been developed over the years for ordinary opamps don't work with this new breed of amplifier.

Voltage feedback opamps have high impedance inputs. Both inputs are open-circuit from an ideal viewpoint. For a current feedback amplifier the inverting (–) input is ideally a short-circuit to a buffered version of the (high-impedance) non-inverting (+) input.

For a voltage feedback opamp wired as an inverting amplifier, band-limiting the signal is done by putting a capacitor from the output to the inverting input. This capacitor can be adjusted to give the desired bandwidth. However, adjusting the capacitor does not affect the LF gain of the circuit at all. This technique requires that the amplifier is *unity gain stable*.

This shunt capacitor technique cannot be used on current feedback amplifiers. If a capacitor larger than a few picofarads is connected as described, the amplifier will just oscillate furiously. The only way usually mentioned to adjust the bandwidth of a current feedback amplifier is to change the feedback resistor. Unfortunately this also changes the DC gain.

FIGURE 10.8A:



Although this technique was apparently known to some people since 1988,[2] it was sufficiently unknown to be independently invented and published as a new technique in 2000.[3]

Whilst this configuration has a negligible effect on a voltage feedback opamp [(bias current) × resistance × (noise gain)], it has a very strong effect on a current feedback amplifier. The idea is to change the effective resistance from the output to inverting input, without changing the gain of the amplifier. With this SEEKret under your belt {available} you can make current feedback amplifiers do what you want.

Although voltage feedback opamps can be labelled as 'stable for gains greater than 5', for example, this designation is not appropriate for current feedback amplifiers. You could say that all current feedback amplifiers are unity gain stable. However, and it is an

[2] D. Potson, 'Current Feedback Op Amp Applications Circuit Guide', *Application Note OA-07* (Comlinear Corporation, 1988, now part of National Semiconductor Corporation.)

[3] L.O. Green, 'Potentiometer Tames Current-Feedback Op Amp', in *EDN* (Cahners), May 11, 2000, pp. 177.

important point, you must connect at least a certain minimum amount of resistance from the output back to the inverting input to achieve stability. Thus a current feedback amplifier has a ‘minimum feedback resistance required for stability’.

Current feedback amplifiers had a brief period of extra interest around 1990-2000 when their bandwidths exceeded those of voltage feedback amplifiers by perhaps 10×. However voltage feedback amplifier performance then took a massive step forward so that there is again no advantage to current feedback opamps in most applications.

10.9 Imperfections

Don’t expect opamps to do everything wonderfully, regardless of what elementary text books or manufacturers data sheets might suggest. In reality opamps have a large number of imperfections:

- ☹ Current flows into the input terminals (bias current).
- ☹ The bias current has noise.
- ☹ FET input bias current increases strongly with temperature (doubles every 10°C).
- ☹ The bias current can change significantly with common-mode input, and this change is not linear.
- ☹ The bias current into the two inputs is not equal (offset current).
- ☹ The input capacitance is non-linear, and for source impedances >10 kΩ harmonic distortion can be high above a few kilohertz, even for high-impedance FET input opamps.
- ☹ The bias current at low frequencies (less than a few kHz) will have a dominantly 1/f characteristic; flicker noise.
- ☹ The opamp has an input offset voltage, (usually less than 10 mV).
- ☹ The input offset voltage has a TC.
- ☹ The input offset voltage is noisy.
- ☹ The input offset voltage noise at frequencies below a few tens of hertz will have a dominantly 1/f characteristic (flicker noise).
- ☹ The gain at DC is not infinite.
- ☹ The gain-bandwidth product is not infinite.
- ☹ The open-loop gain response has at least two poles.
- ☹ The output is harmonically distorted because of the non-linearity of the output stage.
- ☹ The harmonic distortion increases with signal amplitude.
- ☹ The harmonic distortion increases with load current.
- ☹ The harmonic distortion increases with frequency.
- ☹ The output has a finite *slew rate*, defined in V/µs.
- ☹ The output impedance is non-zero.
- ☹ The output is not able to swing to both power rails when loaded, and for older bipolar designs may only swing to within 2 V of the power rails.
- ☹ The common-mode input range does not necessarily extend to both power rails.
- ☹ Large inputs can cause output phase-inversion in some opamps.
- ☹ Some opamp inputs can be blown up by differential input signals larger than ±0.7 V.
- ☹ A common-mode input produces an output (finite CMRR).

☹ The output changes due to fluctuations on the power rails (finite PSRR).
☹ The maximum output current may be less than 30 mA, depending on the exact type chosen.
☹ Higher output loads cause heating of the output stage and this couples to the input stage causing a long thermal tail {a slow drift of input offset voltage}.
☹ Clamp diodes across the input pins can cause the input impedance to be non-linear at RF and microwave frequencies well outside the bandwidth of the opamp, resulting in DC offsets when large out-of-band RF interference is present.
☹ Higher speed opamps often only have ±5V power capability or even just +5V.
☹ Opamps powered from rails larger than ±18V are rare.

This is a lot of imperfections for the perfect device presented in elementary texts. You should also be very careful when using manufacturers' macro-model simulations of opamps. Do not assume that the opamp simulation is a true representation of what the opamp will do in the real world. Typical problem areas where the macro-model may not be representative of the real device include:

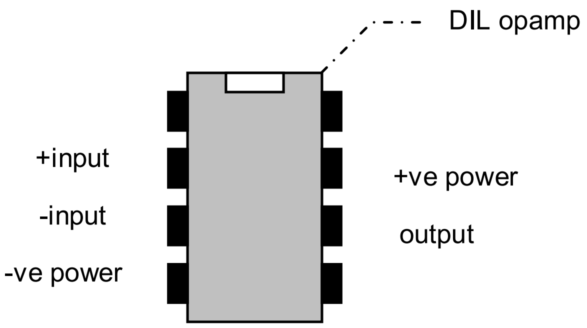
☹ Overload recovery simulation
☹ Power supply rejection
☹ Bootstrapped power rails
☹ Non-linearity of pulse response with signal swing.

It may be necessary to draw out the macro-model from the net-list given to see just what has been modelled. If, for example, there are current sources, resistors or capacitors connected to node 0, the signal ground, this model will not be useful for simulating bootstrapped power rails.

On one design of 12-bit ADC buffer, I was measuring the system ENOB (*Effective Number Of Bits*) at various frequencies when I noticed that the ENOB had dropped from 11.4 bits to around 10 bits as a result of removing an emitter follower buffering the opamp output. This discrete circuitry was within the feedback loop of the opamp to maintain DC accuracy.

When ENOB drops by 1 bit, the performance of the amplifier in terms of noise and/or distortion has got worse by a factor of ×2. Hence the emitter follower and current source that I had included, at a component cost of perhaps $0.10, had more than *doubled* the dynamic performance of the overall system. Needless to say these parts were immediately replaced.

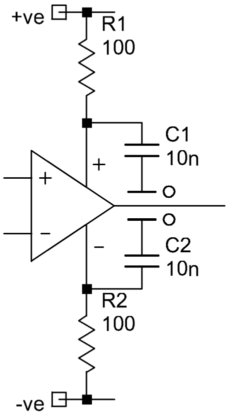
FIGURE 10.9A:



***EX 10.9.1:** An opamp is connected as an inverting amplifier with a gain of 100 and an input resistance of 100 kΩ. There is 0.8 pF capacitance between adjacent pins due to the package and lead-frame. There is a sinusoidal noise voltage on both power rails of 10 mV ptp at 45 kHz due to a switched-mode power supply. What level of signal at the output might be expected due to this power supply noise (neglecting other imperfections in the opamp). Hint: what pin is the –ve power pin next to?

Putting the negative power pin next to the inverting input has another drawback in addition to the capacitive coupling just examined. Current into the power pin can couple via mutual inductance into the inverting input pin. Opamp designs, c.2004, changed the package layout [†] to eliminate this effect. The inputs are all on the left, the power and outputs are on the right.

FIGURE 10.9B:



It is easy for the internal power supply rejection ratio to be worse than the capacitive coupling mechanism; it depends on how good the opamp is. It is also quite usual for the PSRR to be considerably worse (>10 dB worse) on one of the power rails. The cure for both problems is the same however. Just put a small resistor in series with the power pin and then decouple it to ground with a capacitor. Typical resistor values range from 10 Ω to 1000 Ω and typical capacitors from 1 nF to 10 μF, depending on the frequency of the noise sources that exist in the system.

Application notes show the capacitors but not the resistors. The application note is trying to show you how few external components are necessary to make the circuit work. This has more to do with marketing than engineering. The capacitors have little impedance ‘to work against’ and therefore provide little benefit. Also, power supply noise gets coupled straight into the nearby ground tracks, possibly creating a worse noise situation.

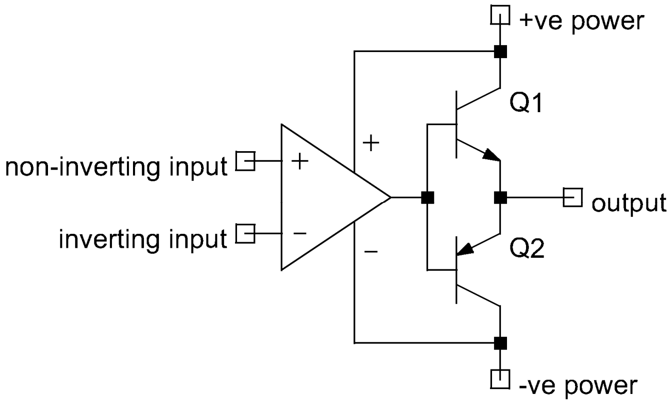
There are three situations where you might use ferrite beads instead of the resistors, or use nothing at all:

- The opamp is a heavy duty type supplying hundreds of milliamps of load current.
- It is an output stage, and is not sensitive to power supply noise.
- It is such a low cost or small size design that the extra components cannot be tolerated.
- The opamp is part of a dual or quad opamp and the resistors would cause too much ***cross-talk***.

[†] Analog Devices AD8099 for example.

Best practice is to use these power rail filter components unless you have a good reason not to. These networks are best not shared with other opamps, though, especially in a multiple stage amplifier. The cross-talk via the power rails can cause oscillations in a high gain chain of amplifiers, or a lack of channel-to-channel isolation in a multi-channel amplifier system. In any case, using the same decoupling network for multiple opamps spread across a circuit board makes the routing poor and the decoupling less effective.

FIGURE 10.9C:



This equivalent circuit for an opamp demonstrates a key point about where the output current of an opamp comes from. If you think of an opamp using this model you will never be confused as to the power source and signal current flow. And don't be distracted by the fact that my opamp model contains an opamp!

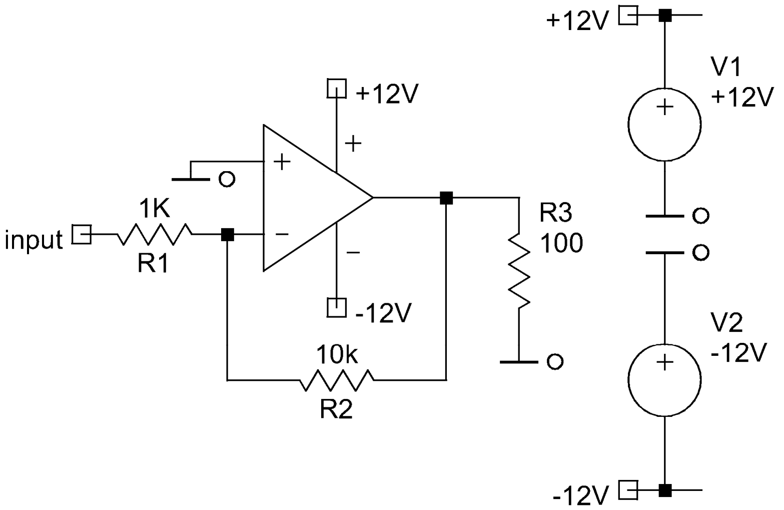
The model has two parts: an output stage and everything else. The output transistors have very high current gain, requiring minimal current from the amplifier stage.

Current flowing out of the *output* terminal is seen to be coming from the +ve power rail through Q1. Current going into the output is seen to be going through Q2 to the –ve power rail. This simple model and an exercise further demonstrate the need for power supply decoupling.

FIGURE 10.9D:

***EX 10.9.2**: On the circuit shown, a step input signal is causing the output to slew upwards (positively) from zero volts.

Draw the *complete* load current path.

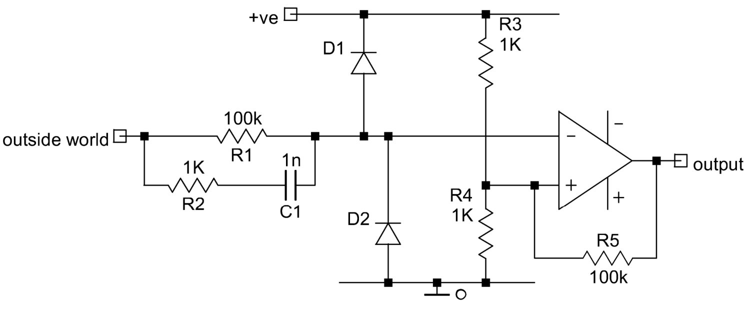


10.10 The Schmitt Trigger

A ***Schmitt trigger***[4] takes an analog input signal and converts it into a digital high or low state. It is a 1 bit ADC with hysteresis. Schmitt input digital gates are also useful for eliminating noise problems from digital input lines. When interfacing to switches, for example, it is wise to use Schmitt input gates whenever possible.

Microprocessor based systems do not require Schmitt gates; the switches are generally read twice, with a delay of a few tens of milliseconds in between, thereby eliminating noise and contact bounce problems.

FIGURE 10.10A:



In this circuit R1, R2, C1, D1 & D2 highlight the fact that you should *never* connect an unprotected circuit to the 'outside world'. The circuit will get destroyed by static electricity or by *homo stupidus*[†] wiring it up to a power rail.

The correct device to use for a Schmitt trigger is a comparator not an opamp. A comparator is specified for use in the *saturated* output condition necessary for this circuit. However, a low speed circuit (<10 kHz) will often function to some extent with an opamp instead. The only reasons to do this are cost or PCB space; dual or quad opamps can be bought for little more than the cost of singles. Thus using a spare opamp might save $0.10.

When using high speed comparators (< 2 ns propagation delay) you should be aware that a greater amount of overdrive will slightly change the propagation delay. Thus there will be a certain amount of jitter on the output signal due to the rate of change of the input signal. This propagation delay dispersal will only be a problem if you are worried about propagation delay variations of less than a few hundred picoseconds.

It is very unwise to use a comparator without using hysteresis (positive feedback). A comparator is a very high gain amplifier. At the desired switching point, any noise at the input is heavily amplified and this can cause the output to thrash up and down between its maximum and minimum output levels. Typically the amount of feedback should exceed the peak noise at the input by at least 10 mV in order to avoid switching problems. Typical feedback networks use both a resistor for the low frequency feedback and a small (pF range) capacitor to speed up the edge.

In the above circuit, the switching point is set by the ratio of R3 to R4 and, to a lesser extent, by R5. The hysteresis is set by the ratio of R5 to the parallel resistance of R3//R4. Thus there is some slight interaction between the values; changing R3 to change the ratio, and hence the switching threshold, also changes the combined resistance and hence the hysteresis. This interaction could be minimised by making R3 & R4 small, then putting a series resistor to the comparator input. This new series resistor should be used

[4] O.H. Schmitt, 'A Thermionic Trigger', in *Journal of Scientific Instruments*, XV (1938), pp. 24-26.

[†] Latin sounding name for a stupid user.

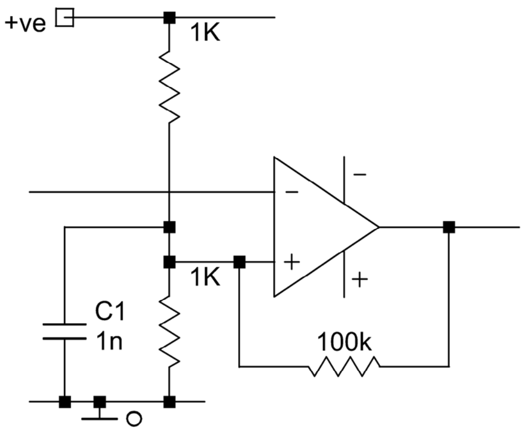
to dominate the Thévenin source resistance of the network; changing R3 or R4 would not then dramatically affect the hysteresis. It is really a question of how much you intend to keep changing the values as to whether the additional cost of this extra component is justified. Ordinarily I would think not.

Remember that the cost of adding a resistor is not just the $0.005 of its material cost. You also have to take into account:

☹ Increased circuit diagram complexity.
☹ The lower reliability of an increased component count.
☹ Increased PCB area used, possibly either making the board bigger or making the tracking more difficult.
☹ The extra work of preparing the bill of materials.
☹ Increased time to populate the PCB.
☹ Increased PCB layout effort.
☹ More soldered joints means more cost on a hand soldered board.
☹ Another component value on the automatic insertion program.

Each of these effects could be almost vanishingly small on its own, but together they make a reasonable argument for not unnecessarily adding components. The overall impact of these effects depends on how the board is being assembled, whether or not the board is tight on space, the total number of components on the board &c. Just be aware that a component costs more than its ‘face value’, and each $0.005 resistor added may well have turned into a $0.10 cost increase on a finished unit. Some of the boards I have designed had between 1000 and 3000 components on them. Adding an extra 5% of components at $0.10 each would not have made me popular!

FIGURE 10.10B:



***EX 10.10.1:** Your new trainee, *Junior*, has come up with a plan to prevent power supply noise from affecting the threshold of the Schmitt trigger. He has decoupled the threshold point with a small capacitor, C1. Will this work?

***EX 10.10.2:** Your new trainee, *Junior*, now proposes to cost-reduce the overall circuit by removing the inverter following the output of the Schmitt trigger. To keep the signal phase the same, the comparator inputs need to be swapped over, whilst still putting the feedback to the positive input. He has just proposed this at the lunch table, so there is no circuit to look at. What is your response?

CH11: the ADC and DAC

11.1 Introduction to Converters

Some engineers are never quite sure whether ADCs/DACs belong to the analog domain or the digital domain; this is all a question of application. A 12-bit converter never gives 12-bit performance. There is firstly the consideration of noise. A large amount of noise will swamp the quantisation levels so that resolution down to 1 LSB is not possible. Then there is the question of dynamic performance. Static performance measures are DC gain and DC linearity. Dynamic performance measures indicate what really happens when you apply a signal, although the tests are often done with a sine wave.

Effective Number Of Bits, ENOB [pronounced E´-nobb] indicates how well the whole circuit performs relative to an ideal system. If only 5 effective bits are needed from an 8-bit converter, leave it to the digital guys. If better than 7 effective bits performance is required from an 8-bit converter, that's where the analog department is required!

Getting good performance from an ADC or DAC is not something you should take for granted. It takes great skill to get the performance specified in the manufacturer's data sheets, and much of this comes down to the PCB layout. Remember that the manufacturer makes up an isolated test board with clean, noise-free power rails. When you put the device in your system, there will be microprocessors, switched-mode power supplies, motors, quantum dehumidifiers † … all sorts of junk conspiring to ruin the performance of your analog sub-system.

Oh, and by the way, you have a 10-channel acquisition system to make but the application notes talk about *single-point grounding*. You now have to be a lot smarter than the designers of the manufacturer's test boards if you want to get anywhere near as good a performance as they claim.

The very first thing you have to ensure with any ADC or DAC is that the digital guys have done their jobs properly. Check that the levels and edges on the clocks are correct according to the device data sheet. When you get 'random glitches' out of an ADC, perhaps on less than one in ten thousand acquired data points, your first thought should be that the clocks are wrong, or that the data is being latched at the wrong time. It is true that some converters produce spurious results, but on a long established device, the fault is clearly down to your design.

Clocks have to be really good to get acquisitions with no glitches over say 10^9 acquired data points. If the levels are noisy, as seen on a scope, then the 5σ peaks of the noise may cause a problem which you won't explicitly see on a scope. The clocks need to meet the required logic levels with plenty of noise margin to spare.

It is your job to keep pushing the digital guys to get clean, noise-free, jitter-free clocks. If the converter has a differential clock input then you absolutely have to use it differentially, rather than single-endedly, in order to get the best performance out of the device. Put in a miniature RF transformer if you need to, but get a differential clock drive.

The other thing you may need to do is guide the digital guys in their clock

† Fictional noisy device.

distribution path. They will want the clock to be from some $1 phase-locked loop chip, or they will pass an initially stable clock through all sorts of multiplexors, digital frequency dividers, programmable logic arrays &c. All of these devices will contribute to the clock jitter, making the resultant clock of very poor quality. You need to minimise the number of devices that the clock passes through and the aforementioned devices must be considered with respect to the amount of harm they will do to the clock.

Sub-harmonic modulation is an easy problem to spot since it can be seen on a spectrum analyser. Distortion in the spectrum analyser itself produces harmonics, but never sub-harmonics. Therefore any sub-harmonics seen are real. Sub-harmonic modulation can occur, for example, when divided-down clocks pass through the same buffer as the main clock; ***cross-talk*** in the buffer produces the modulation. The simple act of connecting a divider onto a clock signal will inject a certain amount of sub-harmonic modulation back onto the clock line; the changing currents within the divider will feed back to its own input.

Random noise on the clock is much more difficult to measure, but it results in jitter of the acquired (ADC) or output (DAC) signal. The easiest way to investigate these problems is to apply a pure sinusoidal input signal and view an FFT of the acquired data, using at least 4096 points for the FFT in order to get sufficient frequency resolution. The width of the fundamental, provided you use a good ***windowing function***, will indicate whether or not you have excessive clock jitter.

If you are running the clock at 100 kHz on an 8-bit system and you are not interested in the resulting jitter, then don't bother with the previous paragraphs. Otherwise, do a few sums to see how much the clock jitter will affect the performance of your system. In practice, even jitter of a few tens of picoseconds (RMS) can have a strong adverse affect on the performance of a system.

$$\boxed{ENOB \leq\ -0.3 - 3.3 \cdot \log_{10}\left(2\pi f_{SIGNAL} \cdot \delta t_{RMS}\right)}$$

… where δt_{RMS} is the combined RMS value of the sampling clock jitter and ***aperture jitter*** within the ADC.

11.2 Non-Linear DACs

As a system block, a DAC is very simple. Digital inputs are applied and an analog output appears. This is the very simplest sort of DAC and could be just a resistive network. Various DAC versions are available, some of which contain internal latches for the data and others of which allow the data to be applied serially.

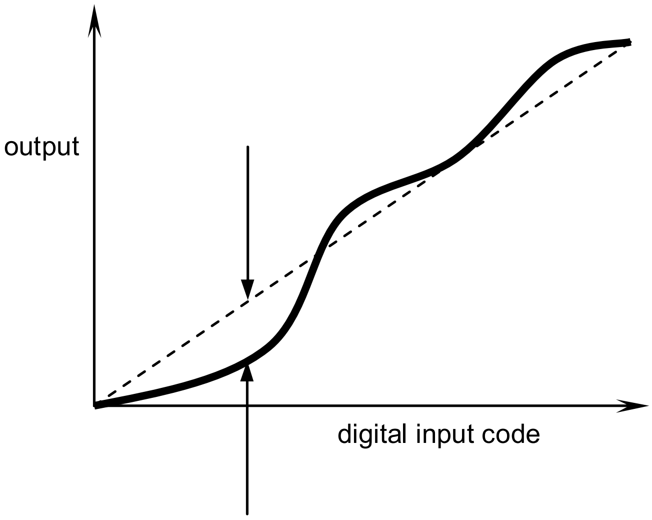
For a latched variety, digital inputs are applied; when they are settled, a clock line is used to tell the DAC to produce an output. The clock and data then have to meet the ***setup and hold times*** of the DAC.

It is a common mistake to fail to check these critical timings and then wonder why spurious results are occasionally produced. It is essential on any new design that the setup and hold times are measured sufficiently accurately to prove that the device will function reliably. It is a necessary *but not sufficient* condition that the prototype circuit seems to work. I don't care if you have also tried the prototype over temperature and over supply voltage variations. It is essential that the timings are verified using a scope. Failure to do so is asking for trouble.

Now I am not talking about any specific technology of DAC here. This is just generic data on DACs. After all, DAC technology is changing very rapidly so technology-based statements may be out of date before this book is published. Serial DACs, for example, are now small and inexpensive. Given the minimal number of pins required, the timing requirements are easy to verify.

As an overview of a high resolution DAC I can draw a picture of its behaviour. This transfer function curve is grossly exaggerated in order to show the errors clearly. For the purpose of illustration, the DAC steps have been made so fine that they cannot be seen on this scale. What is clear is that there is some sort of non-linearity in the output waveform. It is necessary to quantify this error in order to compare one DAC against another. *Non-linearity* as a term is not sufficiently explicit. There are several ways of defining the non-linearity, and it is important to understand the differences between these definitions.

FIGURE 11.2A:



One of the simplest linearity definitions is found by drawing a straight line between the end points of the transfer characteristic and measuring the deviation from this line in terms of LSBs. The dotted line between the end points is the linearity reference. The worst linearity error is the biggest vertical line between the reference and the actual transfer characteristic. When looking at the overall transfer function of the DAC, the term used is *Integral Non-Linearity* (***INL***); when measuring relative to the end-point reference line this becomes *end-point INL.*

For the transfer curve shown, you should be able to visualise a better reference line that could be fitted to the transfer curve; a *best-fit* straight line. Given that it is a best-fit, the measured INL will be less. Thus if a manufacturer wants to make his ADC *seem* better, he may well give you best-fit INL data. Depending on the exact transfer curve of the device, the best-fit INL figure could be twice as good as the end-point INL figure!

It is then necessary to mathematically define what is meant by “best fit”. A ‘traditional’ best fit line would be a ***least-squares*** fit, this form of curve fitting being attributable to Gauss. In other words you minimise the sum of the squares of the error between the best fit line and the actual transfer response. All of this mathematical manipulation is categorised under the heading of *linear regression.*

As far as the manufacturer is concerned, a least-squares fit will not give the lowest INL error for the data sheet. The lowest INL will result from making the regression line as a *minimax* fit, this form of curve fitting being attributable to Tschebyscheff. The minimax method minimises the error magnitude. What then happens is that there will be at least two points on the transfer function where the maximum error occurs; one being a positive error and the other being a negative error. These will also be the INL maximum values. Although software is readily available to compute least squares linear regression lines in programs like Excel and Mathcad, you may still need to write your own iterative loop to locate the best fit line in your own verification software.

The other major non-linearity term for a DAC is the *Differential Non-Linearity*, DNL. Looking in detail at the individual LSB steps, it is seen that they are not all of the same size. In fact it is possible that increasing the digital code by one LSB might actually decrease the output (when an increase was expected). In this case the DNL is so bad that the DAC has become ***non-monotonic***.

Each 1 LSB step on the digital input should give rise to a 1 LSB change at the output. Gain and offset errors are ignored by comparing the output step to the size of a mean LSB output step; DNL is measured relative to this mean step. –0.7 LSB DNL means that one or more steps are only 0.3 LSB in amplitude. +0.9 LSB DNL means that one or more steps are 1.9 LSB in amplitude. Often converter DNL specs are given as ±0.7 LSB, for example, but there is no need for the positive and negative figures to be the same. A DAC with a DNL spec of –0.9 LSB and +1.2 LSB is still monotonic.

11.3 DAC Glitches

Suppose you actually tried to plot the DAC curve given earlier. If your test software went systematically through each code, you might be expecting the output voltage to move smoothly from one voltage to the next after some sort of settling delay. What actually happens is that the output changes in some horrible noisy manner, eventually settling to the final value. The horrible noisy transient is the ***glitch.***

If the manufacturer's part has horrible glitches on its output, this fact will not be mentioned in the data sheet. To find out how big the glitches can be, you may have to look at another DAC which has a better performance. This better DAC may be described as a 'low glitch' part and may show curves of its performance against a cheaper non-low-glitch type. Marketing people shout about their product's good spec points and don't mention the bad spec points. Since the glitch amplitude and duration are both necessary to determine the adverse effect, you will find glitches specified in terms of *glitch impulse*. The glitch impulse is defined as the area under the curve of voltage against time.

Glitch impulses have historically been huge and external glitch reduction circuitry has been necessary. Fortunately, modern DAC designs should not need active external glitch reduction circuitry. You should expect to have to use a simple passive filter however. To give an idea of the order of magnitude of glitch impulse that is produced, the 10-bit LTC1663 plot of its midscale glitch has an estimated area of around 60 nV·s. The 16-bit LTC1650 quotes a "low glitch impulse" of 2 nV·s and the 14-bit DAC904 quotes a glitch impulse of 3 pV·s. Notice that the glitch impulse magnitude from one type of DAC to another can be several orders of magnitude apart and therefore any old DAC will not always do the job. As newer parts are developed, you should expect to see a reduction of these glitch impulse figures on the more expensive parts.

Consider the case of the 16-bit LTC1650. Its output glitch is around 10 mV, which for a 4 V FS output represents a 164 LSB deviation! The glitch response of a DAC is therefore vitally important in any non-static situation.

The uses of DACs fall quite neatly into two distinct categories: static and dynamic. The static case would be for something like an automated adjustment, emulating a gain or offset pot for example. In this case the glitch response and settling time will probably not be too critical.

The dynamic application is a considerably more stringent requirement. A typical dynamic situation is where the DAC is being used to reconstruct a waveform. Suppose you synthesise a sinewave using the DAC. Any glitches on the output will degrade all the dynamic measures of this sinewave, examples of such dynamic measures being THD+N, SFDR and ENOB, all of which are explained in the appendix. If reconstruction is what you are doing, get a DAC that has these dynamic figures given or you will get caught out.

Glitch impulse specs are usually given as typical for one point on the scale. It is up to you to prove that the glitch impulse does not get worse at other points on the scale. The glitch impulse will almost certainly vary with code output and not necessarily in a ***monotonic*** fashion.

Seeing a typical glitch impulse quoted on a data sheet, your first question should be: "how much does it vary"? The answer can only really come from the manufacturer. If glitch impulse is important to you, hassle the manufacturer by phone, fax, and email. This is where buying power becomes important. If you are a small manufacturer and you want 100 pieces a year, the manufacturer may not be keen to answer your questions. If on the other hand you do $20,000,000 a year total business with this manufacturer, they may be more forthcoming with data. In any case you must always try asking.

The problem for you, the outsider, is that you don't know what mechanisms cause the glitch impulse spec in this particular device. If it is a simple capacitive effect you might assume a tolerance of say ±15%. The glitch may correspond to a coupling capacitance between the control line and the output. Such a glitch would be fairly well defined by the process.

EX 11.3.1: A 5 V control line on a DAC internally couples to the output via a parasitic capacitance of 1 pF. The DAC output impedance at this point is 10 Ω. What is the maximum possible glitch impulse that will result?

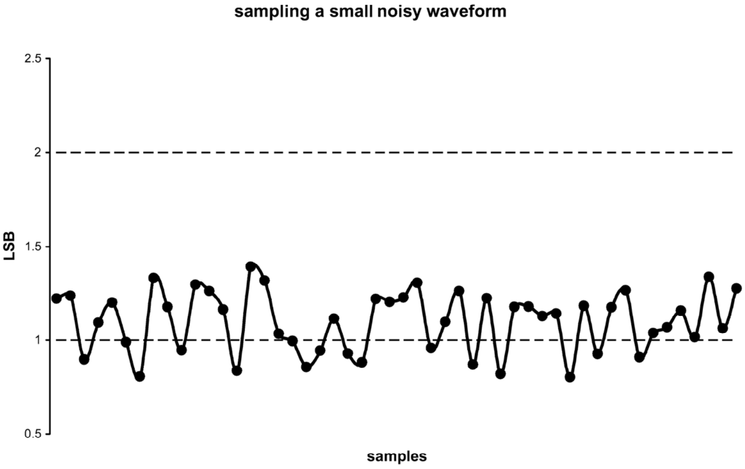
The problem is that you do not know if the glitch being produced is being reduced by balancing one effect against another. A typical trick to minimise the output glitch would be to use identical switching transistors driven from anti-phase clocks, giving first order cancellation of the glitch. Now, a slight change in one device will have an exaggerated effect in the final result. This makes it impossible for you to estimate how much variation can be expected in this particular spec. If you can't get any more information from the manufacturer, your only recourse is to test several sample devices from different batches to get some sort of estimated variance.

11.4 Dithering An ADC

In English the word 'dither' means a tremble, quiver or vibration. In electronics it is a small deliberately injected signal used to enhance the accuracy of a sampling system. Dithering is widely applicable to many differing situations and adds great value to a sampling system. It is best illustrated by example.

If you look closely at an ADC at the LSB level then you can see the conversion process in action.

FIGURE 11.4A:



The horizontal lines represent the start and end of the sampling levels. A data point between the 1.0 and 2.0 lines will be recorded as a 1.

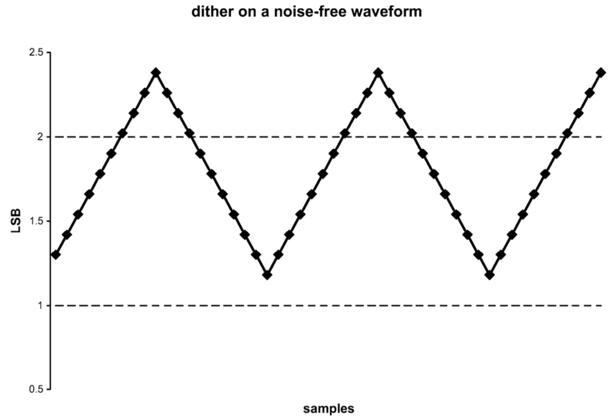
If the waveform shown is sampled and the resulting data averaged, the value will be somewhere between 0 and 1, depending on exactly how many points are above the 1 line and how many are below. The mean value will therefore have more resolution than the LSB quantisation level.

Because the noise is less than 1 LSB ptp the waveform might sit entirely within one quantisation band and there would be no information available as to its position within the band. Thus the available resolution will drop according to the position of the noise band relative to the conversion levels. In this case, just adding random noise to the system to make the total exceed 1 LSB ptp, then averaging, will result in an acquired signal that does genuinely have more resolution, regardless of the actual position of the signal within the quantisation band.

Increasing the noise to just over 1 LSB ptp and averaging does enhance the acquisition performance, but that is not the best solution. The added noise is reduced by the averaging process, but some remains. If the noise is Gaussian, one should expect the noise power to reduce directly by the number of points averaged. Thus the noise voltage reduces as the square root of the number of points averaged. A better solution, therefore, is to add a small defined waveform to the signal. If the mean value of this added signal is deliberately made zero over the averaging period, no noise will have been added and a resolution enhancement will still have been achieved.

This added signal is called a *dither* waveform. Common shapes used for dither waveforms include triangular and sinusoidal, the additional improvement achieved using more complicated dither waveforms being debateable. The triangular dither waveform is the easiest to understand.

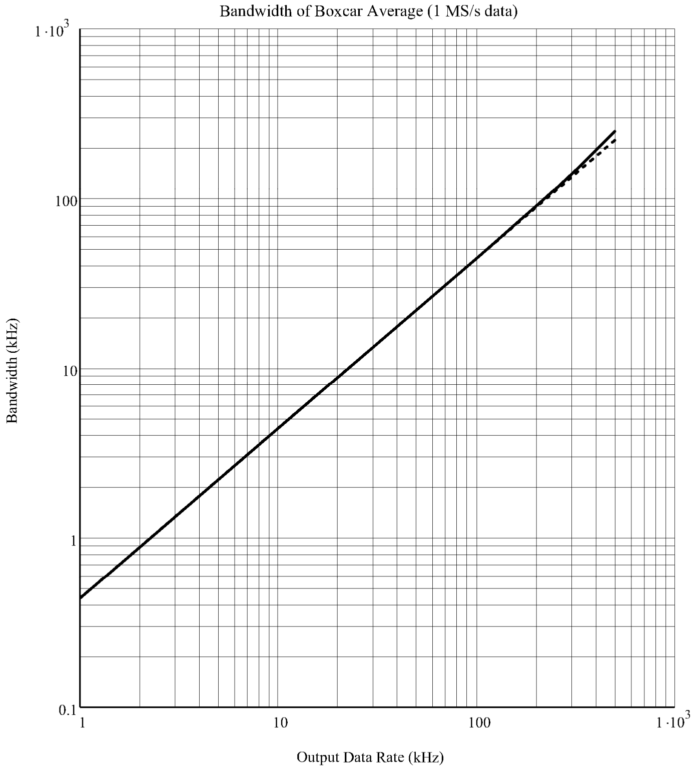
FIGURE 11.4B:



The dither waveform needs to be at least 1 LSB ptp for this scheme to work. The mean value of the dither waveform is zero if the sampled data contains an integer number of these triangles. It is preferable that there is only one cycle of the dither waveform per group of acquired points; this means that any level on the dither waveform is only sampled once. This system is known as an *over-sampling and dithering scheme*.

The ADC is sampling at perhaps 100× the output data rate in order to get enhanced resolution. A group of sequential data points is digitally filtered to produce a single data point with enhanced resolution. A simple way of doing the filtering is to just sum the data points together. Notice that this scheme averages the dither signal to zero, but it also puts a low-pass filter on the acquired data as well.

FIGURE 11.4C:



The exact equation for the bandwidth of a block accumulated scheme, also known as a *boxcar average*, is quite unpleasant (see appendix). However, the approximation, shown dotted, is simple and accurate:

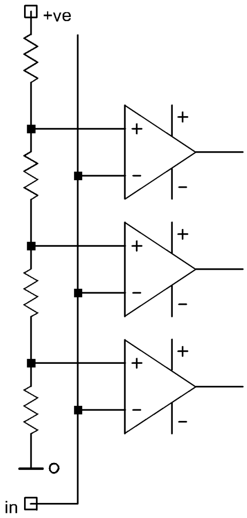
$$\boxed{B \approx 0.443 \times F_{OUT}}$$

If a 100 MS/s data stream is averaged in blocks of 10 then the output data rate, F_{OUT} , is 10 MS/s and the bandwidth is reduced to 4.43 MHz. This formula does not work for a block size of 1, ie no averaging, as the bandwidth would not be limited by the 'filter'. The formula and graph rely on the actual system bandwidth being much greater than the predicted filter bandwidth. If the system bandwidth is more than 10× the filter bandwidth, the error caused by neglecting the system bandwidth will be less than 0.5%.

This ADC enhancement scheme looks too good to be true, and it is. You will not be able to take an 8-bit converter and enhance its ENOB from say 7 bits to 9 bits using this scheme. The reason is that this dithering scheme is ***interpolating*** between the LSB thresholds. As such, it does not correct the inherent DNL and INL of the converter. Therefore the result is not very accurate and less ENOB enhancement is achieved than might be expected.

For use on integrating ADCs, where the DNL is very always very good, the 1 LSB dithering scheme is excellent. However, it is not as useful for flash converters, where the DNL can easily be as bad as ±0.9 LSB. In this case the DNL of the converter also has to be decreased if a good increase in ENOB is to be achieved.

FIGURE 11.4D:



Here is a reminder of where the DNL error of a flash converter comes from. The LSB steps are generated by a resistive ladder. Ideally the steps would all be exactly equal, but in reality the resistor tolerances and comparator input-offset voltages create differences.

In addition to the random errors, there may be a systematic error in some of the levels if you start looking below the 1 LSB level. To see what sort of magnitude of error is involved, take the example of an ADC with a 2 V reference. For an 8-bit system, 256 levels over 2 V= 7.8 mV per step. Given that untrimmed comparator offsets can be around this sort of value, it is easy to understand why the DNL specs can be so bad. To get better DNL specs on fast ADCs, active trimming techniques are used on the unpackaged silicon chip.

In practice, DNL is the major contributor to errors in flash converters and if it is corrected, greatly improved resolution, accuracy and ENOB are achievable using oversampling.

The solution is to spread the dither signal over several quantisation levels, perhaps 4 or 5. In this case the individual DNL errors average out and the resultant is greatly enhanced. As an example, this technique has been successfully used to enhance the accuracy of an 8-bit converter from an ENOB of around 7 bits to an ENOB of 10 bits.[1] Remember that ENOB includes noise and distortion, so an extra 3 bits (8× improvement) is a considerable enhancement. This is one of those intellectual property items where the cost of getting the enhancement is a development cost and not a unit cost increase. The additional component cost per scope was around $3 which was negligible on a product selling for $7000.

It is clear that dithering and averaging reconfigures an N-bit acquisition system into an (N+E)-bit system at a lower acquisition speed, where E is the number of enhanced bits achieved. If M is the number of over-sampled points used to produce one output point, then it is easy to see that for a 1 LSB ptp triangular dither waveform, the resolution enhancement is M extra steps within each LSB. Thus $E \le \log_2(M)$. However, when dithering over several bit thresholds in order to reduce the DNL errors, the DNL errors behave more like random noise. In this case the resolution enhancement is only by the square root of M. $E \le \log_2\left(\sqrt{M}\right)$.

Using the rules of logs $E \le \frac{1}{2} \cdot \log_2(M) = \frac{\log_{10}(M)}{2 \cdot \log_{10}(2)}$ or $\boxed{E \le 1.66 \times \log_{10}(M)}$

It is unfortunate that manufacturers have been quoting resolution enhancement by just averaging for many years. You see stupid numbers like 13-bits resolution from an 8-bit ADC system. Engineers realise that this is just marketing gone mad and neglect it. On the other hand, if you take an 8-bit system that is capable of dithered over-sampling and apply a 0.5 LSB ptp sine wave, you will see nothing much with the trace magnified to say 1 div ptp if the over-sampling is turned off. When the over-sampling is turned on and you suddenly see a really clean sine wave, only then will you truly believe this is real.

[1] Gould Instrument Systems, model Classic 6100: released c.1998.

The system just described does oversampling and dithering. This reduces the bandwidth of the measurement system. However, it is widely known that just averaging multiple sets of measurements of the same signal reduces the noise and therefore improves the signal-to-noise ratio. It changes the bandwidth of the noise, but does not affect the steady part of the signal at all. Dithering can be used between successive measurement groups. These groups can then be averaged to give enhanced accuracy and resolution, but without changing the signal bandwidth at all.

It is not necessary to make the dither steps finer than 0.01 LSB. In fact 0.05 LSB steps would probably also be acceptable. The reason is that it is unlikely for the system noise to be less than 0.1 LSB, so the unintentional noise will effectively add greater resolution to the dither signal anyway.

11.5 Characterising ADCs

You will see data sheets from manufacturers showing graphs of DNL and INL that show resolutions well below the LSB level. You might think that such a plot was achieved by stepping the input up gradually using a precision calibrator and measuring the exact switching points. This method is possible for 8-bit converters with 1 MHz analog bandwidth, but ADCs are available with greater than 20-bits resolution. Also ADCs with 8-bit resolution can have analog bandwidths in excess of 2 GHz. With these specs there is no chance of measuring the devices in wafer test jigs or packaged part jigs using the simple DC method. In any case, the ADC itself may have more than 1 LSB of peak-to-peak noise.

For these reasons, manufacturers effectively use over-sampling techniques to get the precise INL and DNL readings for their devices. For example, feeding a slow ramp into the ADC and averaging data points can give 100 or more points over the span of one mean quantisation level. This is really no different to using a triangular dither waveform and gives higher resolution to interpolate the exact quantisation levels. Manufacturer's data sheets for ADCs can therefore appear misleading because they give DNL curves with resolution down below the level of 0.1 LSB. This performance is not achievable unless you are using an oversampling scheme.

Since the input data is a ramp there is not too much uncertainty in the resulting data. There will also be some noise on the waveform and in the conversion system and this may cause the ADC reading to jump to an adjacent level. If you think in terms of ADC bins, as one does when doing histogram plots, then noise can cause the ADC output to jump to a bin at least one step away from the correct bin. Suppose that due to the present level of the ramp, the output should be in bin 40. A few values may also fall into bins 39 and 41. However, when the ramp level is at 39 or 41, noise may cause some of the readings to fall into bin 40. Thus on average the noise will not affect the histogram.

A mean of 10 hits per nominal bin is too few to guarantee the result. Suppose this is an 8-bit system. That means 256 levels or bins. If 2560 points are acquired, there is an expectation of 10 hits per bin if the ramp starts at the beginning of the sweep and ends at the end of the acquired data. The ramp should overscan the ADC range slightly, in order to guarantee to get all the ADC levels.

If the converter has a DNL of ±0.8 LSB then you should expect one or more bins containing only two points. It is possible that noise will make no readings fall within the bin on the first acquisition sweep. The ADC would then be declared faulty because of a missing code. One hit in a bin would make the DNL read as –0.9 LSB. Two hits in a bin would make the DNL read as –0.8 LSB. Depending on the application, at least 30 hits per bin, on average, is desirable in order to make incorrect fault reports less likely, and ≥100 hits per nominal bin is preferable in order to get reasonable resolution on the result.. If there are 30 hits per nominal bin, a –0.8 LSB DNL bin would therefore get 6 hits on average.

30 hits per nominal bin is a tough requirement on a 16-bit converter. There are 65536 bins, and therefore 1,966,080 points or more are needed. One might then take lots of smaller acquisitions in order to fill up all the bins.

This ramp technique, using over 200 nominal hits per bin (a 1Mbyte store), has been used routinely on production calibration tests of 12-bit scopes[†] to guarantee that their DNL is not worse than ±0.8 LSB.

EX 11.5.1: [special interest only] With 30 hits per nominal bin, *estimate* the probability that a ±0.8 LSB DNL ADC will be reported as having a missing code in the histogram test given above.

Another use of this ramp/histogram method is to re-map the ADC levels, in other words to correct the linearity of the ADC. Low cost ADCs can have very poor linearity, easily seen on the ***FFT*** of a pure sinewave. This poor performance can be corrected by the ramp/histogram method, the incoming 8-bit values being re-mapped to more accurate 12-bit values using a lookup table.

The lookup table is created from the histogram, assuming that the ramp is totally linear. Starting at one end of the ADC range, the next output position is determined by the number of hits in the appropriate histogram bin, divided by the mean number of hits expected in the bin. This simple technique has been used in production 8-bit DSOs, improving the SFDR from worse than 48 dBc to better than 68 dBc.[‡] Note that the test ramp was specified as being adequately linear, but there was no way to verify the spec. However the fact that the SFDR improved when the ramp data was used to recalibrate the ADC levels proved that the ramp was in fact adequately linear!

As introduced earlier, ENOB is a quantitative measure of an acquisition system's performance. An ideal *N*-bit ADC will have an **E**ffective **N**umber **O**f **B**its of *N*. Noise, DNL, INL and dynamic distortion in the ADC, and any buffer amplifier, will reduce this performance. Manufacturers of ADCs often quote ENOB values for their converters at various input frequencies. Alternatively the manufacturer may quote SINAD or THD+N, all of which are interconvertible measures. These provide a quantitative way of comparing one manufacturer's part against another.

It is very easy to get good error performance at low speed (<1000th the quoted analog bandwidth of the device), but as the signal frequency is increased, clock jitter becomes more significant. This is clock jitter in both the user's applied clock signal and the clock distribution path within the device. For this reason alone '8-bit' ADCs can drop to 6 or 5 effective bits at their quoted maximum bandwidths.

ENOB is a sinewave test of the conversion system. The formula used being:

[†] Nicolet Accura 100, released late in 2001.

[‡] Nicolet Sigma 60, released mid 2003.

$$\boxed{ENOB = \frac{1}{6.02} \times \left[SINAD_{dB} - 1.76 + 20 \cdot \log_{10} \left(\frac{\text{FULL - SCALE AMPLITUDE}}{\text{ACTUAL INPUT AMPLITUDE}} \right) \right]}$$

This formula is derived in the Appendix.

EX 11.5.2:

a) A signal generator has a THD+N spec of 0.01%. Comment on its suitability for testing the ENOB of a 12-bit ADC system.

b) A signal generator has an SFDR of 60 dBc. Could this be used to check the ENOB of an 8-bit ADC?

Always use the best generator/oscillator you can lay your hands on when doing an ENOB test. If you are looking for SFDR values better than 60 dBc above 100 kHz available signal generators may not be good enough. In this case a narrow band filter can be used to remove noise and distortion from the signal source.

A generator with a digital display will have noise associated with the display and is therefore probably more noisy than a plain generator. A multifunction generator will also be potentially more noisy than a single-function generator. The best possible signal source is therefore likely to be an ‘old fashioned’ single-function generator, rather than an all new, flashy multi-function, “all singing, all dancing” version.

The generator output always has noise over a wide range of frequencies and may also include specific spurious frequencies as well. These emissions are only going to degrade the measurement, so it is sensible to filter them out. What you want is a passive tuned-LC filter with a modest Q, say >10. A low-pass filter with a cut-off above the frequency you are measuring is another option, although this is not nearly such an effective filtering strategy.

The response of the filter is important. It mustn’t introduce distortion, otherwise the THD+N of the generator could be made worse. Thus you have to be careful about the choice of inductor if making your own filter. Ideally the filter should attenuate by 20 dB or more at a factor of 2 from the centre frequency [Q > 7]. The filter will then considerably reduce any second harmonic distortion in the generator, as well as limiting the bandwidth to other noise sources. Air-wound inductors are preferable, being inherently linear.

Now you can acquire some data and feed it into a computer for analysis. From the definition of SINAD, it can be seen that you need to find the mean-squared difference between the acquired waveform and the best fit sinewave. It is not a trivial matter to get the best fit sinewave, as it is necessary to adjust the amplitude, phase, frequency and offset of a roughly correct sinewave to get it to be the least-squares best fit. It is then easy to get the mean squared difference, evaluated over an integer number of cycles for best accuracy.

SINAD can also be evaluated by doing an FFT of the acquired data and RSS summing everything other than the fundamental. The problem comes in terms of the ***windowing function*** used in the FFT and the resulting finite width of the fundamental. This makes the FFT approach quick and approximate, but not definitive. If you wish to measure SINAD using an FFT then use a long FFT, certainly not shorter than 4096 points.

If you are measuring SINAD on an 8-bit converter then almost any window function (not rectangular) on the FFT will be acceptable. The only time you could (and should) use a rectangular window is when a whole number of cycles of the pure sinewave fit exactly within the FFT length; ideally this whole number of cycles should be a prime number in order to get as many different (vertical amplitude) sampling levels as possible. This unusual and contrived situation means having a pure oscillator phase-locked to the acquisition clock. It is the method used by ADC manufacturers to test their devices and explains why the fundamental is always only one bin wide in their FFT plots.

On a 12-bit (or more) ADC it is essential that you use a high performance window function. A Blackman-Harris, windowed-Sinc, or better is recommended. Having taken the FFT you then need to discard the points around the fundamental and at DC. Take the remaining data points and sum the squares of the actual values, not the dB scaled values.

Having obtained an ENOB value, I am sure that you will be able to improve the circuit to get a better value. This improvement will follow a law of diminishing returns; getting the last few parts of an LSB improvement will be the hardest.

As an aid to this debugging process, you can average 100 or so complete data records. These need to be well aligned in terms of starting at the same point on the waveform. By averaging them, you will minimise the noise. Hence the ENOB will improve and the remaining errors will be due to non-linearities in the converter and the amplifier. This gives you a clue about what you need to improve. Alternatively, if you are able to do an FFT on the data points, you will immediately be able to see harmonic distortion and spurious noise sources.

Averaging 100 or more complete waveforms will also have the effect of oversampling the data. Don’t be surprised if you end up with more effective bits than the nominal converter can give. An 8-bit converter may give an ENOB greater than 9 bits using this method.

Distortion on an amplifier output may be due to a low value of load resistor and the inability of an opamp to drive that much current. One solution is to use a larger load resistor if possible. Another solution is to use a better amplifier or possibly to add an output buffer stage. However, a more subtle form of distortion can occur if a stage is oscillating. Such an oscillation may not be visible in the output data because of the overall system bandwidth.

This is a real-life example: A 200 MHz switched-gain amplifier was being optionally filtered by a single LC stage at 1 MHz. When the 1 MHz filter was selected, the FFT of the data showed 57 dBc SFDR due to harmonic distortion. This was surprising because the unfiltered system showed more like 80 dBc SFDR. The harmonic distortion occurred equally for input signals over a range of 1 kHz to 100 kHz. Now it is usual for a higher frequency signal to be more distorted than a lower frequency signal; at lower frequencies there is more loop-gain available to reduce the distortion. Constant distortion with frequency is unusual, and confusing.

These facts suggested that the inductor in the filter was non-linear, and yet it had previously been fine. Then it was found that this poor SFDR only occurred on certain gain settings of the stage driving the 1 MHz filter. But then again this switched gain stage was perfectly linear (good SFDR) in the 200 MHz bandwidth mode, that is when the 1 MHz filter was not selected.

Unfortunately this sort of interaction is not uncommon in analog design and it can certainly cause some head scratching. You try to follow the logic of this through and you end up with illogical answers. The inductor looks non-linear, but didn't seem that way earlier, and only seems non-linear on certain ranges; that can't be the problem. The switched gain amplifier is linear without the filter, but not linear on all settings with the filter; the amplifier can't be the problem. The low frequency load of the filter on the switched gain amplifier is the same as the load when the filter is not selected, so the load can't be the problem. What does that leave, … nothing!

Sherlock Holmes, the fictional expert detective, is often quoted as having said something like, "When you have eliminated the impossible, whatever is left, however improbable, must be the truth." Well that is great, but what *usually* happens is that you eliminate everything and nothing is left!

When everything has been eliminated, you have to assume that an earlier eliminated item was discarded prematurely. Sometimes a colleague or boss can come along at this point and ask all sorts of 'dumb' questions which cause you to re-evaluate some prior discarded possibility and you can then find the fault.

In this particular case, however, I picked up a scope probe and found to my horror that the switched gain stage was oscillating at around 200 MHz, only when the 1 MHz filter was selected, and only in one range of the switched gain stage. The 200 MHz oscillation was not seen because the two-pole 1 MHz filter killed it off effectively, but that did not stop some sort of modulation effect giving rise to the distortion.

Although the filter load was resistive at low frequency, it was obviously reactive at high frequency and the switched-gain amplifier didn't "like" this load in one of its ranges. It would probably have been possible to kill the oscillation by placing an additional load resistor at the input to the 1 MHz filter. The problem with this would have been that the resistor would have needed to be around 50 Ω or lower, and this in itself could have made the distortion worse. The final solution was to put a small capacitor in series with the extra load resistor. The amplifier then only "sees" this higher load at some high frequency, outside the passband of the 1 MHz filter. Effectively the amplifier gain is reduced at the highest frequencies, without introducing a phase shift, and this makes the stage more stable. This sort of series RC ***snubber*** circuit is often used to damp out problem areas.

@EX 11.5.3:

a) Why does a low value resistive load increase harmonic distortion?
b) How can a low value (shunt) load resistor increase the stability of an amplifier?

There are many lessons to learn from this experience. One is that when eliminating possible sources of a problem, you have to be very sure that you have eliminated them. Cut the track to that stage, pull out all the critical components, pull out the whole PCB; whatever you do, be sure that whatever it was cannot possibly be causing the problem now. When you *are* sure, you can move on. When there is a nagging doubt in your mind, it slows you down.

***EX 11.5.4:** An 8-bit ADC has a 2 V FS input range and an actual DNL of ±0.75 LSB. The combined noise of the ADC itself and the signal conditioning system gives an effective RMS input noise of 100 μV.

a) Estimate the worst PTP noise for 100,000 data points.
b) Estimate the worst RMS noise for 100,000 data points.

11.6 PCB Layout Rules

Another aid in the noise/ENOB/SFDR debugging process is to look at the output data when the input is slowly ramping up. The transitions from one bit to the next should be clean. If there is feedback from the output data lines to the analog input then you can find that the major transitions of the converter are particularly noisy. Consider the converter at midscale. The binary output code changes from 0111,1111 to 1000,0000; this amount of data-lines changing simultaneously gives the maximum possible feedback to the input.

Actually the manufacturer's data sheets and application notes are very helpful in this respect. The manufacturer has had to lay out this particular converter and make it work well in order to tell you how wonderful it is on the data sheets. You would do well to listen to their proven advice.

One standard rule is not to run the digital outputs under or near the analog inputs. This reduces possible input to output coupling. Another rule is not to run the digital outputs directly under the IC package. This is a pretty harsh rule because it is often convenient to route tracks under the package to make the tracking easier. It is a good rule, however, because otherwise there is crosstalk from the digital output lines to the semiconductor die, giving the same problems as running the outputs near the input. If you put a ground plane on the component side of the PCB under the ADC this crosstalk problem is removed.

Remember, if you lay out the PCB with tracks under the ADC and it gives any sort of digital output to analog input feedback, you will not be able to *prove* that this is the fault until you re-lay the PCB. ADC layouts are not very forgiving and if you need to move ground and power tracks, you may well find that it is not possible to physically model the change effectively on the old version of PCB.

You then get into hoping that the new layout works, and perhaps waiting a few weeks to get the new prototype board back in your hands – only to find that it still doesn't work! This sort of iterative design work chews up project timescales and gives you a bad name. The other side of this is that if you always lay the board out correctly first time, then nobody knows what a genius you are, and how difficult it was to get the layout correct! It seems that the only way to show what a genius you are is to take an existing layout that *someone else* has been struggling with for many months, re-lay it and have it work first time. *Then* you get the praise that is your due!

Another rule is to not have the ADC outputs changed by external sources during a conversion cycle. If the ADC has tri-state outputs and they are connected to a digital bus, then the bus has to be "quiet" during the acquisition cycle. If not then there can be crosstalk from output to input again. You may well find that an additional external tri-state buffer is a more cost-effective solution to a noise problem than trying to get the system to work without.

The grounding of ADCs is an area of great debate and 'expertise'. There are all sorts of wise words spoken about it and you have to decide which advice is the best to follow, since they are not always in agreement. Generally the ADC has at least two ground pins, and frequently it has three or more. These pins have interesting names such as analog ground, digital ground, output ground &c. These are all 0 V connections, but they have been separated within the device to reduce internal *ground bounce* due to the finite impedance of the bond wires. Often the best solution for the interconnect seems to be to join them all together into a copper area directly underneath the IC body. By doing it this way there are no via holes to add (undesirable) inductance.

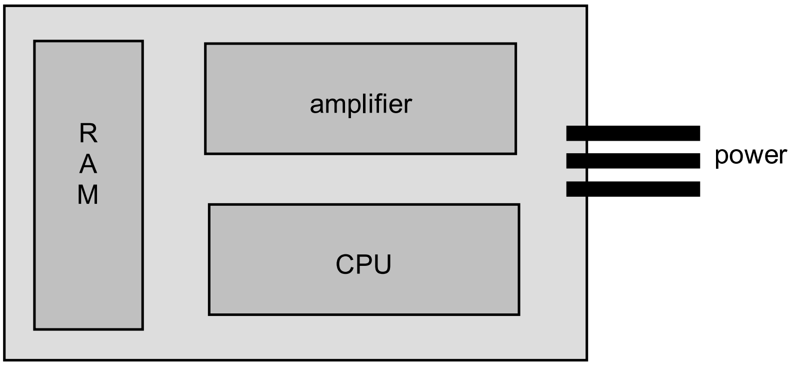
As for what you then do with this little copper land {area}, what I would say is that you should leave your options open. If you are not sure about cutting (or linking) the ground plane in a particular area, then cut it and put links on an exposed face of the PCB so that you can try re-making the cut if there is a problem. Give yourself a *fall-back position* if your original idea doesn't work as well as intended.

Try to figure out where the currents are going to flow and discourage noisy digital currents from flowing through your nice sensitive amplifier. Remember that a current has to both go and return, so think about the whole current path. If a digital line goes into an analog area then its (noisy) return current will have to flow through the quiet area. Maybe you need a small RC filter on this digital line on the way into the analog area to keep the noisy HF current out.

The various circuit areas of a PCB or a system need to be segregated {separated} so that noisy currents don't flow through sensitive areas. This is an essential part of system design. After you have a circuit diagram, you can put mental or even dotted boxes around parts of the circuit that are sensitive and that should be kept clear of noisy circuitry. By 'noisy circuitry' all I mean is circuitry that has current or voltage transitions of any appreciable size or frequency. Digital circuitry is noisy when it is changing state or being clocked. This noise is not only on the signal lines, but also on the power and ground wires/planes.

Here is a typical lousy PCB design.

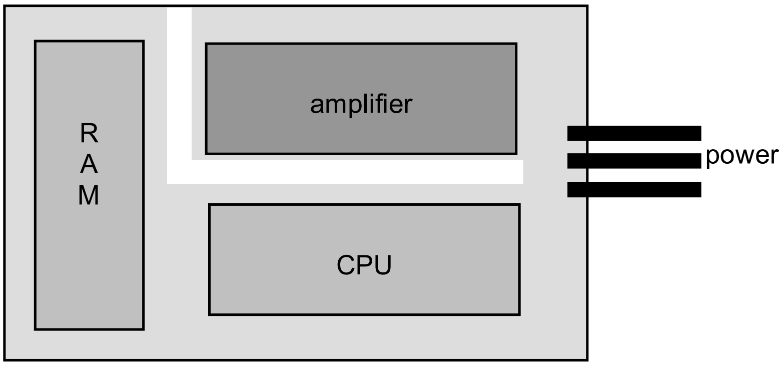
FIGURE 11.6A:



If this layout is on a solid ground plane then the noisy RAM and CPU currents will flow through the amplifier. Fixing this, without repositioning all of the components, requires selective cutting of the ground plane.

This cut-out is illustrated below.

FIGURE 11.6B:



You must also prevent tracks from being routed through the amplifier on their way to somewhere else. This can be achieved by putting a 'routing keep-out' on all layers of the board. Routing keep-outs like this may not be convenient or popular, but they do work. The amplifier might require a simple screen over it (emphasised by shading it differently above). This screen could be photo-chemically machined {chemically milled} out of brass, copper or tin-plate very cheaply ($1)

When fitting a screen over components on a PCB, it is vital that the screen does not lay directly on top of the PCB tracks {traces}. *Solder-resist is not a guaranteed insulator*. Solder-resist stops solder sticking to the tracks and prevents minor contaminants, such as dust, bridging between tracks. It will not stop a metal screen shorting to an underlying track, and it will not stop an inadequately coated resistor end-cap resistor shorting to an underlying track.[†] PCB mounted screens need to be shaped {profiled} so that they sit at least 0.5 mm above the PCB where tracks run under the screen.

Ground plane noise is usually impossible to eliminate on an existing PCB. The board needs to be re-layed in order to fix the problem. The difficulty the designer has when faced with a non-working board is therefore to decide, and if possible to prove, that the noise is unfixable without a board re-layout. If the noise was from some other source, changing the layout will not have the desired effect.

Typically ground plane noise could be from a few microvolts to a few millivolts referred to the input of the amplifying devices. This level of noise is extremely difficult to measure directly. An incorrect direct measurement technique is to try to measure the difference in voltage between two points on the ground plane using a scope and a standard 10:1 scope probe. The 10:1 probe makes the scope too insensitive to measure the ground noise.

A 1:1 scope probe is also of little use. The bandwidth achievable through a 1:1 probe is not more than a few megahertz and its ≈60 pF load capacitance will slug any signal it is connected to anyway. It will not be adequate to effectively measure clock and data noise from a digital system. Remember that a ground plane is a very good short-circuit at DC, but becomes progressively less of a short-circuit as the frequency increases, by virtue of its inductance.

Interestingly, if you do manage to measure noise that appears to be in the kilohertz region using a DSO, the frequency is probably an ***alias***! Test for this by applying a 1 MHz bandwidth limit in the DSO, for example. If the noise drops dramatically then you know that the frequency was an alias.

The best way of measuring ground plane noise is with a differential probe. You can

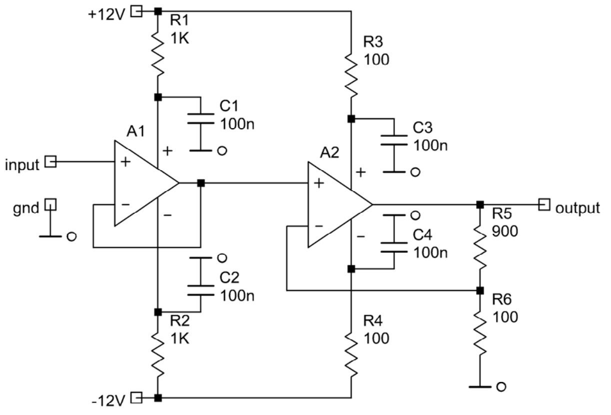
[†] This has happened and in one particular case, c. 1998, it cost tens of thousands of dollars in re-work and recall expenses.

make you own using a miniature RF transformer and a piece of coaxial cable. The RF transformer will give no response at DC, but then again it is only the changing signal which is likely to cause a problem. You will need to use a small piece of PCB material to make a suitable mount for the RF transformer and to provide a mounting point for the coax output lead. The absolute calibration of the probe is not as important as the ability to measure some sort of signal. In this case it is likely that a spectrum analyser will give a better display of the ground plane noise than a scope.

If there is high frequency current flowing through the ground plane then there will be significant volt drops between different locations on the plane. ‘Significant’ here still means less than a millivolt. Now this level of signal in the ground plane will not capacitively couple to a wire to any great extent. Therefore if you think that an opamp ground reference, for example, is not seeing the correct signal due to a volt drop across a ground plane, then it is easy enough to lift that pin and route a separate wire as a test. It is not unreasonable to have a relatively clean ground plane and then to have a few key ground tracks that are not connected to the ground plane. They may need to form their own separate ground path back to the input of the amplifier stage, for example. Fortunately this is now easy to model because you just link the sensitive pins together with a wire, and this is effectively what you will be doing in the next iteration of the PCB as well. However, it is possible that the ground plane may couple by mutual inductance to this new ground wire. Thus the final ‘isolated’ zero volt signal track is best kept as far from the ground plane as possible.

The issue of which connections should go on this ‘select’ ground path is not something that can be written down easily. One has to understand the current paths, rather than follow memorised rules.

FIGURE 11.6C:



This circuit encapsulates many design concepts given to this point in the book. It is a real circuit for a high input impedance 25 MHz ×10 amplifier, using appropriate values and techniques for the function performed. A1 is a FET input opamp and A2 is a bipolar input opamp.

***EX 11.6.1:**

a) Suggest why R1 and R2 are ten times bigger than R3 and R4. Is there any significance to this 10:1 ratio?
b) R5 and R6 give quite a significant load to A2. Can they safely be increased by a factor of say ten?
c) What is the point of A1? Why can’t A1 be given gain or A2’s input be the main input?
d) Given that the best available opamps are already being used, how can the performance be further improved?

***EX 11.6.2:** In the above circuit the power rails are quiet and all the 0 V connections are made to a solid ground plane. Even so there is still too much noise on the output. This noise appears to be random digital noise, presumably from the digital circuitry on the same board. The ground plane is only connected to analog circuitry and yet it seems to be 'infected' with digital noise. The digital power and ground planes have been routed specifically to avoid being adjacent to this ground plane so that capacitive coupling from them to this ground plane should be minimal. A senior expert maven guru consultant has come up with some waffle about needing to re-lay everything from first principles. His mantra {mystic chant} is, "See the current: be the current." Can you suggest anything more immediate to prove the point?

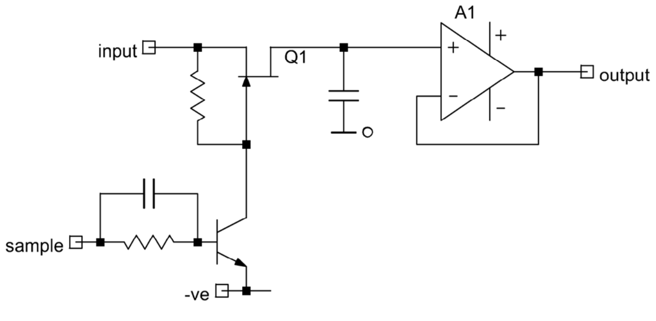
11.7 The Sample & Hold

If you want to sample for 1 ns, and hold for 1 s [as an analog system], you cannot do it with a single sample & hold. You would need < 100 pF for the 1 ns sample and > 100 nF for the 1 s hold. In this case the answer is to use a *dual-rank* sample & hold. The first samples in 1 ns, and holds for say 30 µs. The next can sample for 30 µs and hold for 1 second. This system design has taken us from a $\frac{\text{hold time}}{\text{sample time}}$ ratio of 1,000,000,000 to a ratio of 33,333. An impossible design has become an achievable design.

This brings up another system consideration. For the previous example, it would have been better to sample for a short period and then use an ADC to convert the value to a digital form. Once digital, the value could be held indefinitely without degradation. The problem is that the analog circuit may well be able to hold the value for 1 s with a defined error. But what about if the receiver of this signal can effectively take the value at any arbitrary time in a 1 ms to 1 s interval? Your boss may tell you to hold the signal for 1 s, but not mention the fact that it will be used after 1 ms in some cases. The inevitable *droop* on the waveform coupled with the variation of time when the sample is accepted effectively gives noise on the output value. The key to accuracy is to force the output value to be sampled after a defined period [if possible]. This technique minimises the noise on the reading.

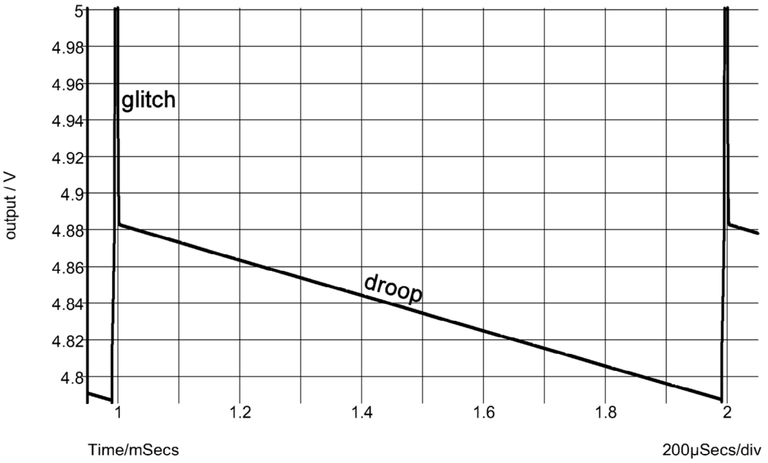
This is a typical discrete component low-speed sample & hold circuit. You have to deal with the ON-resistance of Q1, the off-state leakage in Q1, the bias current in A1, the *charge injection* when Q1 is switched and the layout of the components.

FIGURE 11.7A:



You will save time and money using an integrated solution to this problem, rather than making your own from 'first principles'. If you buy a packaged solution then it will be characterised and you will be able to see immediately if it is going to work. When you make your own, even with a circuit as simple as this one, little circuit quirks {oddities} can arise which take time to fix.

FIGURE 11.7B:



This plot shows two of the problems mentioned. The droop is caused by leakage current and will be temperature dependant. The charge injection glitch is caused by the switching edge on the gate-drain capacitance for the JFET. Notice that the simulation shows all this 'noise', and yet it is sampling a DC signal!

Increasing the capacitance will minimise the droop and the charge injection error. However, the bandwidth will be restricted since the ON-resistance of the JFET combined with the hold capacitor form a low-pass filter.

Given the ever-reducing cost of integrated circuit solutions, the modern method would be to digitise the signal immediately using an ADC locally timed relative to the sample clock. The digitised data could then be read from the ADC by the main processor any arbitrary period of time later, without droop being a factor.

****

Here is an example of an engineering compromise. This book was practically finished. The index was just about complete, and then I realised that the important definition below had been omitted. The book was painstakingly manually indexed and to insert the definition in the section on digital storage oscilloscopes, or in the glossary, would have destroyed at least a week's work and delayed the production schedule for the book. This gave three choices:

☹ Add it in correctly and re-index the book at horrendous cost.
☹ Leave this important information out completely.
☹ Add the definition in a non-optimum position, but where it wouldn't require re-indexing. Put a link in the index to it, hoping the reader would look for it if they didn't know what it meant.

In this sort of situation there is no "right answer". One answer is perhaps less bad than others.

ROLL MODE: At slow timebase speeds on a digital storage oscilloscope (DSO), it is unhelpful to wait for the trace to complete the sweep before seeing anything. Suppose you are on 1 s/div. This requires 10 seconds to see what happened. For this application, some DSO's can be switched into roll mode (typically slower than 100 ms/div). The trace now appears to be rolling from right to left across the screen, much like an old chart recorder. Now the new data immediately appears on the screen. Check for the presence of roll-mode on your new scope purchases. Even very expensive scopes can omit the feature and therefore be very awkward to use for simple tasks.

As a historical note the first modern DSO was designed in the UK by Gould Electronics in the early 1970's. As the originators of the DSO, all Gould DSOs featured roll mode. (A Nicolet DSO pre-dated the Gould offering by about a year, but was more like a digital recording device than a modern DSO.)

A related scope feature is "slow refresh", a very technical insider scope-designer term. In refresh mode the trace typically updates (is displayed) at the end of an acquisition sweep. At 10 µs/div the update appears instantaneous. At 50 s/div nothing happens for ages unless you have slow refresh, where the trace is displayed immediately as it is acquired; the trace updates left to right across the screen. It requires more software and hardware effort to implement slow refresh as the data is being acquired and displayed simultaneously. Hence, not all manufacturers provide this feature.

CH12: the relay

12.1 Benefits of Relays

It was estimated in 1986 that there were 25 billion relays in use.[1] Relays are evidently widely used and can be very important components in a system. It is therefore worth knowing their SEEKrets.

The relay was invented around 1837 [2] – 1840 [3] for use in telegraph systems and the first impression is that it is an old-fashioned, clumsy and inelegant component. Why anyone should use such an apparently noisy and clumsy device in a modern, hi-tech, elegant design?

- ☺ Can connect circuits together using no control power (latching type or when de-energised) and with negligible insertion loss.
- ☺ Control circuit is *galvanically isolated* from the controlled circuit. Leakage current can therefore be orders of magnitude lower than semiconductor circuits and it doesn't degrade too badly with temperature.
- ☺ Smaller power loss and volt-drop in the controlled circuit than with semiconductors.
- ☺ Capacitance from control circuit to controlled circuit can be lower than semiconductor version. This capacitance can be made arbitrarily low by the use of plastic 'push-rod' contact-actuators.
- ☺ Relay switched gain stages do not suffer from as much non-linearity and bandwidth loss as circuits switched by CMOS MUXes and switches.

Relays are of considerable importance for instrumentation and control purposes. There are only a few basic types:

- ➢ Ordinary mechanical relays.
- ➢ Bistable (latched) relays.
- ➢ Reed-relays (dry reeds).
- ➢ Mercury-wetted reed-relays.

There are, however, a vast number of different varieties of each type of relay. Indeed the ordinary mechanical relay group includes high power relays, often referred to as contactors (con-tact′-ors).

12.2 Contact Types

Relay contacts come in three basic types; normally-open, normally-closed and changeover. The term *normally* here means 'when no power is applied'. By convention

[1] H. Sauer, 'Relay-Evolution', in *Modern Relay Technology*, trans. by Naples, J.G., 2nd edn (Heidelburg: Huethig, 1986), pp. 13-26.

[2] W.F. Cooke, and C. Wheatstone, 'Improvements in Giving Signals and Sounding Alarums in Distant Places by Means of Electric Currents Transmitted through Metallic Circuits', *UK Patent Spec 7390* (UKPO, 1837).

[3] S.F.B. Morse, 'Improvement in the Mode of Communicating Information by Signals by the Application of Electro-Magnetism.', *US Patent 1647* (June 1840).

on circuit diagrams, relay contacts are shown in their de-energised state; the relay is OFF.

Form A:	normally-open
Form B:	normally-closed
Form C:	changeover
Form D:	make-before-break changeover

A fourth type of relay contact is make-before-break changeover, but this is considerably less common than the others.

A catalogue might say a certain relay was of type 2C; this would mean it had two changeover contacts. Alternatively, switch terminology is sometimes used by manufacturers. For example, a single-pole single-throw (SPST) switch would be equivalent to a 1A or indeed a 1B relay. A 1C relay would be equivalent to a single pole double throw (SPDT) switch. The number of *throws* is the number of directions a current can be diverted to. The number of poles is the number of input circuits.

relay	switch
1A, 1B	SPST
1C	SPDT-bbm
1D	SPDT-mbb
2A, 2B	DPST
2C	DPDT-bbm
2D	DPDT-mbb

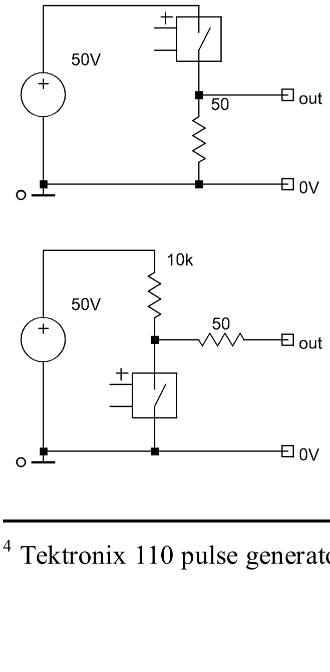
bbm= break-before-make
mbb= make-before-break

form B & form D contacts are rarely used.

Mercury-wetted reed-relays have limited application, given that they contain Mercury (which is poisonous). They do have excellent bounce characteristics, however, so they are useful for making fast risetime high voltage pulses; 50 V pulses with 250 ps risetime, for example.[4] This is not just of historic interest, as a mercury-wetted reed can give an excellent *absolute* standard for pulse response.

One method is to use the relay to switch a resistive load to some known source of voltage, such as a DC calibrator. The 'high' amplitude is therefore known. By leaving the system in this state for some time, any settling delay or warm-up drift can be eliminated. The *return-to-zero edge* can now be viewed. As there is now no source of voltage, there is no problem with source loading or self-heating. The return-to-zero edge is therefore *always* more accurately defined than the driving-to-source edge. This fact is essential when using the edge as an absolute standard of pulse response.

FIGURE 12.2A:



The problem with the method just given is that the pull-down resistor needs to be fairly low in order to get a fast fall-time on the return-to-zero edge. Suppose that a 50 Ω resistor is used: 50 V across a 50 Ω resistor generates 50 W of heat. This requires either a large high-power load resistor or a narrow pulse. The power handling capability of the mercury-wetted reed may also start to be a problem.

Another method of getting a defined return-to-zero pulse is to use the mercury-wetted reed to short-circuit a load resistor to ground. Suppose that a 10 kΩ resistor is fed from a DC calibrator output. If the resistor is shorted to ground by the relay, a low source-impedance fast-edge generator is created.

[4] Tektronix 110 pulse generator (obsolete).

The low level is well defined in terms of the ratio of the ON-resistance of the mercury-wetted reed switch to the 10 kΩ resistor. This technique also gives an absolute standard for pulse response, with the advantage that a higher voltage pulse can be generated than with the pull-up scheme mentioned previously. In this case the reference step response comes from the 'switched to zero' edge.

Although a relay is a near ideal switch in terms of its low on-state resistance and its high off-state isolation, don't think of the switch as perfect. Always consider things like off-state isolation in terms of capacitance if nothing else. The capacitance across open relay contacts can be 0.5 pF or more. Even this small amount of capacitance can be very significant above a few kilohertz when switching-in ×100 attenuators. It can therefore be necessary to improve the off-state isolation by not only opening the signal path, but also shorting the now floating input to ground.

12.3 Reed-Relays

A reed-relay is a small sealed glass tube (20 mm long and 4 mm in diameter for example) containing two or more pieces of ferro-magnetic metal, at least one of which is springy; this glass tube with internal metal contacts is known as the *reed-switch*. The application of an external magnetic field causes the springy part(s) to come together and complete the electrical circuit.[5] Reed-relays typically have an actuating coil wound (coaxially) around the glass tube, this coil carrying between 5 and 50 A·t [Ampère-turns]. Reed switches can be used to switch voltages as high as 10 kV, but in this case they require more like 150 A·t.

The contacts have a voltage rating and a current rating, but the power rating is nowhere near as high as the product of these voltage and current ratings. The Meder KSK-1A69 reed switch, for example, can switch 10,000 V and 3.0 A. However, it can only switch 50 W not 30,000 W. The upper limit for reed switching power is around 60 W.

EX 12.3.1: A reed-relay contains a reed switch with a sensitivity of exactly 20 A·t. The measured coil resistance is 480 Ω. This particular relay actually switches on at 3.5 V at room temperature. Roughly how many turns are there in the relay coil?

The most natural type of reed-relay is a form A type; this also makes it the cheapest and most readily available type. The ferrous material is pulled together by the external magnetic field because this movement improves the magnetic flux path {reduced reluctance}.

Notice that the circuit current flows though the ferrous material and therefore the self-inductance of the relay contact is higher than a similar path in copper wire. Copper plating the ferrous parts helps at megahertz frequencies because the ***skin effect*** makes the current flow through the copper plating rather than through the ferrous material.

A form C contact {changeover} in a reed is not particularly good, even though they are made. Don't even consider the concept of a form B reed-relay. Placing two or more reeds inside the same actuating coil gives a multi-pole relay, reducing the cost and actuating power. Unfortunately there is then a risk of ***cross-talk*** between the different circuits.

[5] W.B. Ellwood, 'Improvements in or Relating to Electromagnetically Operated Electric Switches', *UK Patent Spec 522,798* (USPO, 1938: UKPO, 1940).

Reed-relays can be supplied with an electrostatic screen between the coil and the contact. This screen reduces noise coupling from the coil circuit to the contact circuit via the contact-to-coil capacitance. On inexpensive reed-relays the screen is wound using the same type of wire as used for the coil, a reliable and inexpensive technique. Unfortunately the screen is then an inductor, limiting its effectiveness at frequencies above a few MHz. If this lack of shielding and possible resonance is a problem in your application, specify a copper foil screen, albeit at additional cost.

12.4 Relay Coils

Relay coils are wound of copper; this has a TC of about +0.39%/°C. The TC is important because relays are given a must-operate voltage (also known as the *pick-up voltage* or the *pull-in voltage*) spec at say 20°C. If the relay is being operated in an ambient temperature of 70°C, the must-operate voltage is significantly changed. Remember that the relay is magnetically actuated {activated}. It is normally assumed that the required ampere-turns are constant with temperature. Because the copper winding resistance has increased, more voltage is required to switch the relay ON at higher temperatures.

***EX 12.4.1:** A relay has a must-operate spec of 0.75 of its nominal rated coil voltage at 20°C. What is its must-operate voltage at 70°C?

Temperature rise in a copper winding is also of importance in the safety checking of transformers. If the wire runs too hot, the wire's insulating enamel will break down and the windings will not be electrically stable. For this reason you will find a formula in the safety standard EN60950-1:2006 (Annex E) which is equivalent to:

$$T_{RISE}\,(°\mathrm{C}) = \left(\frac{R_{hot}}{R_{amb}} - 1\right) \cdot \left(234.5 + T_{amb}\right) - \Delta T_{amb}$$

where T_{amb} is measured in °C.

This rule is a good fit to the measured TC of copper over the range of –200°C to +200°C. The ΔT_{amb} term takes account of the fact that the ambient temperature may well have changed in the time that it takes the transformer winding to heat up. Transformers take a long time to reach thermal equilibrium. Even small 250 V·A transformers can take an hour or so to settle down to a steady state.

Check your ohm-meter carefully before doing this test. Some linear ohm-meter circuits 'do not like' measuring the resistance of large inductors. The control loops for the internal current sources can go unstable with the large reactive load.

12.5 Drop-Out

Relays have magnetic parts which come together when the coil is energised. This action dramatically reduces the reluctance of the magnetic path and it is then easier for the parts to stay together. Whilst the must-operate voltage of a relay is typically specified as 0.75 of nominal, the must-release voltage (also known as the ***drop-out*** voltage) is considerably more variable from relay type to type. You might expect a guaranteed drop-out voltage of 0.1 of the nominal operating voltage, for example

This hysteresis in switching points has been used on older designs to provide hysteresis in the controlled system with no additional circuitry. However, such use is not recommended because of the variation in must-operate and must-release voltages. The Matsushita TQ2 relay data sheet, for example, shows variation of must-release voltage from 0.1 to 0.35 of nominal operating voltage over a sample of 50 relays, neglecting temperature effects.

A more modern use for the switching point hysteresis is to reduce the power in the relay. This power saving is achieved by momentarily driving the relay to its nominal voltage then reducing the voltage to say half of that level, thereby reducing coil power by a factor of 4.

For applications where the relay can experience heavy impacts whilst operating, power saving is not a safe technique. A low holding-current might not stop the relay from opening under the stress of a large external shock, and once opened, the contacts may not re-close.

This power saving technique is useful for battery powered applications. It is also useful for precision instrumentation applications. When a relay coil is energised there is extra power being dissipated. This can cause two effects: It can heat up surrounding components and it can heat up the relay contacts themselves.

The ultimate in power saving is achieved by using latching relays. These are available with the same PCB footprint as ordinary relays in the same family. There are two distinct types of latching relays: those with one coil and those with two. For the single coil versions, a “set” pulse is applied in the forward direction. The “reset” pulse is applied in the reverse direction. Since the coil can be driven for less than 10 ms to set or reset the relay, the power consumed is zero for the steady state condition.

If there are several relays to be driven, it is sensible to drive them to their correct state in a sequence rather than to drive all the set/reset pulses at once. This minimises the current transient.

12.6 Thermal problems

In precision instrumentation, it is *vital* that the heat distribution within the instrument remains constant. ***Thermal EMFs*** generated by temperature gradients must be allowed to stabilise and they can then be nulled out. If they change, it is necessary to wait for several minutes, or even tens of minutes, for them to stabilise again before the equipment is operating at full accuracy. Even tilting an instrument at an angle will cause the internal temperature distribution to change. Whether or not this actually causes a change in the instrument’s calibration is a matter for testing.

It is probably fairly safe to tilt a ±0.1% instrument, but I wouldn’t like to guarantee the result with a ±1 ppm resolution instrument; certainly some DMMs would fail this test at the ppm level. If you must move equipment about like this, check it against a known standard before and after movement to see if the calibration has been changed.

Thermal EMFs in relays used to switch low-level signals are a function of the power dissipated in the relay, its internal construction, the contact materials, and also the materials used in the current paths. It is easy to inadvertently create (thermocouple) junctions of dissimilar metals which can generate large (>30 μV) thermal EMFs. Thermal EMFs are not ordinarily a problem for use where resolution on the switched signal is not needed below 300 μV. Thermal EMFs are of vital concern when switching signals with sub-microvolt resolution.

In the measurement of resistors with ultra-precision [<1 ppm], it is usual to have automated test systems that switch-in the reference resistors, switch the leads around, reverse the current direction, allow offset nulling &c. These can then be left running overnight, so that long settling times can be used and multiple readings can be taken. This technique gives a minimum uncertainty on the readings. Such switching systems would certainly use low thermal EMF switching circuits.

It is normal for manufacturers to measure and specify thermal EMFs on reed-relays because they are specifically designed for switching low level signals. Ordinary mechanical relays are often not specified for thermal EMF because they are not rated to switch low level signals. Low thermal EMF reed-relays can easily achieve 10 μV levels [with a time constant on the order of 1 to 5 minutes]. 0.5 μV thermal EMFs are also available.[6] This does not mean that measurements through reed relays are limited to 0.5 μV resolutions, it just means that you need to wait for the thermals to settle out. The thermal EMF then gets subtracted as part of the zero-nulling operations. The requirement is simply that the thermal EMF should not change. If there is a small value to start with, a 1% change is not a disaster.

Thermal EMFs are further reduced by driving the relay coil with two voltage levels as mentioned in the previous section. The relay is first given the full operating voltage until it is firmly on. The applied voltage is then halved, knowing that the holding voltage is usually less than a third of the operating voltage. Halving the applied voltage, quarters the applied power. Note that when there are lots of relays to be driven in this way from an FPGA or microprocessor, the tri-state ability of the logic device can be used to provide all three states from one control pin. For example: hi = turn on, tri-state = hold on, lo=off.

A factor that needs to be taken into account when packing modern miniature relays close together is that they will magnetically interact. The cases can be very thin plastic, with the magnetic paths less than 1 mm from the outside of the component body. Depending on the exact internal magnetic configurations, this *will* change the must-operate and must-release voltages according to the state of the surrounding relays. Relays which are designed to be packed tightly are sometimes characterised for this effect.

If the relay you are using hasn't been characterised for this close packing effect it still suffers from the problem, but nobody could be bothered to quantify it. For Automatic Test Equipment (ATE) jigs it is usual to have hundreds of reed-relays packed at the minimum possible physical separation. To overcome the magnetic interaction problems, the relays often have ***mu-metal*** shields. This shielding has the added benefit of improving the coil sensitivity of the relay (it needs less coil drive power).

12.7 Small-Signal Switching

Reed-relay switches are sealed in an inert atmosphere within a glass tube. Because of the ***hermetic*** seal and the contact materials used (Ruthenium or Rhenium) they can switch circuits with virtually no current and no voltage very easily. These circuit loads are called *dry circuits*.

'Mechanical' relays are *not* rated to switch at zero current and zero voltage. They require some current and/or voltage to break down the contamination films that form on

[6] Coto 3500 series low thermal EMF relays

the contacts.[7] The current/voltage required is very much dependant on the contact material used, the contact force applied and the quality of the seal on the relay (if indeed they are sealed).

When two conductors are touching, there are several factors that determine the resistance of the resulting joint. Obviously the conducting surfaces need to be clean and free from corrosion, but then pressure becomes the key factor. For clean copper conductors the resistance due to the surface contact area can be modelled by the equation:

$$\boxed{R = \frac{23}{A \cdot P^{0.9}}\,\mathrm{m\Omega}}$$

where the area A is in mm^2 and the pressure P is in N/mm^2.

Remembering that pressure is force per unit area, this equation becomes

$$\boxed{R = \frac{23}{A^{0.1} \cdot F^{0.9}}\,\mathrm{m\Omega}}$$

Making the contact area bigger, for a given force, can actually *increase* the contact resistance. In any case, contact pressure of less than 5 N/mm^2 is inadvisable.

Enough pressure will *theoretically* overcome a small amount of surface contamination. In practice, pressure alone is never enough and connections are designed to be made using a slight *wiping* action. By scraping the surfaces together, thin oxide films or organic contamination can be removed, reducing the pressure required to make a low resistance contact.

Relays, mechanical switches, and edge connectors require the surfaces to slide over each other to ensure a good contact. For crimp-on terminals the pressure is so high that a gas-tight joint is formed, preventing degradation due to oxidation.

For equipment in service for long periods, it is sometimes necessary to unplug then re-plug edge connectors to remake the sliding contact. The low pressure contact allows oxidation and contamination of the mating surfaces. This same situation occurs in hand-held infra red remote controls for TVs. Wiggling or rotating the batteries makes an apparently dead controller spring back to life!

In a relay, the contact force is supplied by a magnetic force which is, in turn, supplied by the coil current. This magnetic force is given by: $\boxed{F = \frac{B^2 A}{2\mu_0}}$. This equation gives clues for the effective design of a relay. The magnetic path needs to have a large cross-sectional area, maximising the total flux. However, at the gap it is better to put in a smaller cross-section piece, thereby increasing the flux density.

Although magnetic materials are always non-linear when the field strength gets high, the flux density is roughly proportional to the coil current and hence the coil voltage. Thus for a given design of relay, the contact force doubles for a $\sqrt{2}$ increase in coil voltage. The contact force is therefore proportional to the coil power.

Copper oxidises heavily on contact with air and is therefore unsuitable as a contact

[7] H. Sauer, 'Contact Resistance', in *Modern Relay Technology*, trans. by Naples, J.G., 2nd edn (Heidelburg: Huethig, 1986), pp. 51-57.

material. The ideal properties for contact materials are high conductivity, low oxidation in air (fairly inert), high melting point and great hardness.

Although the *wiping action* of contacts cleans off surface contamination and makes a better connection, this wiping action also limits the lifetime of the contacts if the materials are too soft. Gold, for example, is very soft; whilst it is an excellent conductor, not forming oxide layers easily, it is not ideally suited to electrical contacts in its pure state. Better contacts can be made with alloys of gold. But even a gold alloy contact can be ruined if somebody handles the contact with their (greasy) hands.

Relays can be designed to switch high powers, or small signals, or both. The distinction is often the hardness of the contact material. Gold-plated contacts will wear away too quickly when switching tens of amps. But then again, the materials used for tens of amps will not switch millivolt signals correctly. This problem has been solved by making the contacts layered. It is the contact surfaces that are important. The underlying materials are relatively less important as they are thin and highly conductive.

If the outer layer is gold, the relay will correctly switch low level signals. If the underlying material is hard then should high currents be required, the gold will burn-off very quickly, leaving a hard contact material underneath. This makes a versatile relay, but the user (designer) has to be aware that once the high level currents have been switched for a while, the relay will no longer be capable of switching the low level signals.

A relay designed for switching power circuits may well not specify a minimum switched load. Beware! If the relay data does not state a minimum signal switching level, it just means the manufacturer doesn't intend to sell into that market; if you use such an unspecified relay for small signals then that's your tough luck. Furthermore, if the manufacturer only makes power relays, their technical people may not even realise that there is a problem switching low level signals.

A relay contact will *carry* much greater current than it will switch. Relay manufacturers know this, but never state it. It is therefore possible to use an under-rated relay as an isolating device, but not the main switching device. The relay would still have to be able to break the main current if the primary switch failed, but that would be a single operation cycle, not every cycle.

Making and breaking high currents rapidly burns away the contact material and limits the service life of a relay. By using a semiconductor to do the main switching, the relay could have a much longer life, or a significantly lower current rating. Clearly this depends on the actual construction of the relay, but a factor of two or more on current carrying capability for a relay is not unusual. Going beyond a factor of ten using this technique is definitely pushing your luck {risky}. Test the relay to ensure it can break the current when required. If the relay is not up to the job, the arc will not break quickly, or possibly it will not break at all. Remember that an arc is almost like a short–circuit; whilst the relay is arcing, the load circuit is still 'made' {connected}.

A relay has a significant operating time. This will be variable from device to device and there will be some variation of this operating time during the life of the relay. Big relays with large gaps obviously have a longer operate-time than miniature reed-relays. A reed-relay might have an operate-time of 1 ms. Modern mechanical relays, however, can achieve operate-times in the 1 ms - 10 ms region as well.

12.8 Contact Bounce

Contact bounce is the result of two pieces of metal being thrust together under the influence of either a spring or a magnetic force. The contacts are hard and the spring/magnetic force has no damping. It is an elementary problem in mechanics to see that the contacts will bounce and this can cause considerable trouble to electrical and electronic circuits.

The first major problem has to do with contact wear. When switching a power circuit, you get current through the contacts and voltage across them at the same time during the bounce period. This causes arcing and consequent wearing of the contacts.

Most cars built between 1920 and 1980 had a cam-actuated switch in series with an auto-transformer to generate a high voltage pulse to the spark plugs. A capacitor [in automotive-terminology a 'condenser'] was placed across the switch contacts [the 'points'] to reduce the contact wear. In electrical and electronic systems where power is being switched by a relay you can still find capacitors across the relay contacts. However, it is more usual to put a small resistor [10 Ω - 100 Ω] in series with the capacitor to create a ***snubber*** circuit. The capacitor on its own would create the additional problem of an unlimited switch-on current into the relay contacts from the capacitor.

Switches connected to electronic systems generally need to be *de-bounced*. The switch contacts still bounce, but the electronic circuitry does not 'see' the bouncing. On a large piece of equipment, with a lot of front panel switches, hardware is often not needed for switch de-bouncing. The switches will almost certainly be arranged in a *switch matrix* and scanned. The software will read the keys periodically to see if anything has changed. If so, it then does key debounce by reading the matrix several times until it gets a consistent answer.

Here is a tip for matrix keys in sensitive analog equipment. Don't scan the keyboard all the time. Arrange all the matrix strobe lines to be active at once and read the whole key matrix in one go. This tells the system that *a key* has been pressed. Once a key has been detected, the system can scan the matrix to find out *which key* during the debounce time. This saves on processor time and generates less RFI because the matrix strobe lines are inactive until a key is pressed. The system still generates noise when a key is pressed, but disturbance on the readings when pressing keys is probably not as important anyway.

This is a broad tip for all analog electronic systems. When performing a measurement, shut down as much of the digital activity as possible. This will make it easier to get a low-noise performance. On one system I heard of, the main microprocessor clock was stopped for a short while to allow a low-noise measurement to be made!

In order to give some comparison of analog debounce techniques, I have done a simulation. This pulse train is too regular for a real switch, but it gives a good feel for the problem.

FIGURE 12.8A:

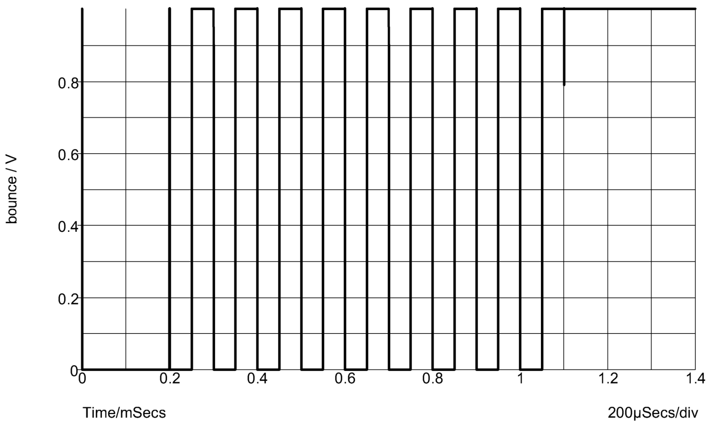
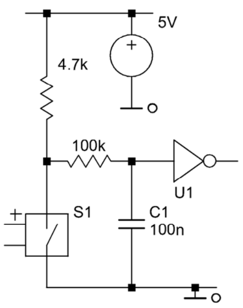


FIGURE 12.8B:



S1 represents the bouncing relay or switch contact. It is vital that U1 has a Schmitt input, otherwise none of these debounce schemes will work reliably.

This is the long-time constant approach to debounce. It is ok for slowly operating switches, but it is no good for switches that need a fast response time.

In order to demonstrate the nature of the circuit, I have reduced C1 to 5 nF for the simulation below.

The non-monotonic response shows the need for a Schmitt input gate. Notice that the switching point of the Schmitt gate may be at 1/3rd of the power rail, this value occurring roughly 1.1 time constants after the edge. Such a delay is not a problem for most applications, but operation at >10 cycles per second would not be possible.

FIGURE 12.8C:

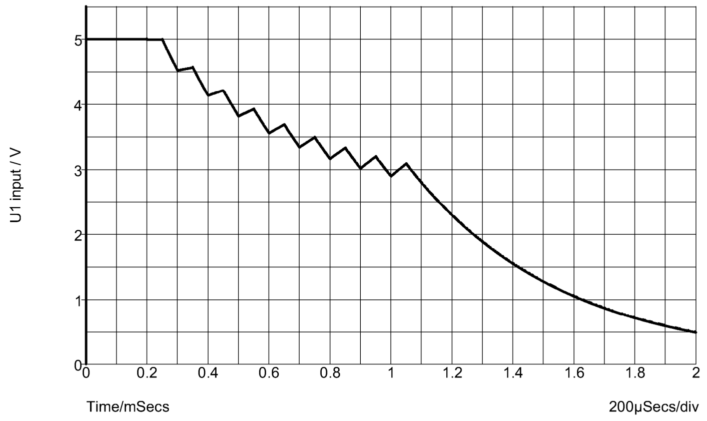
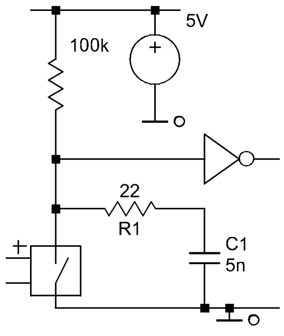


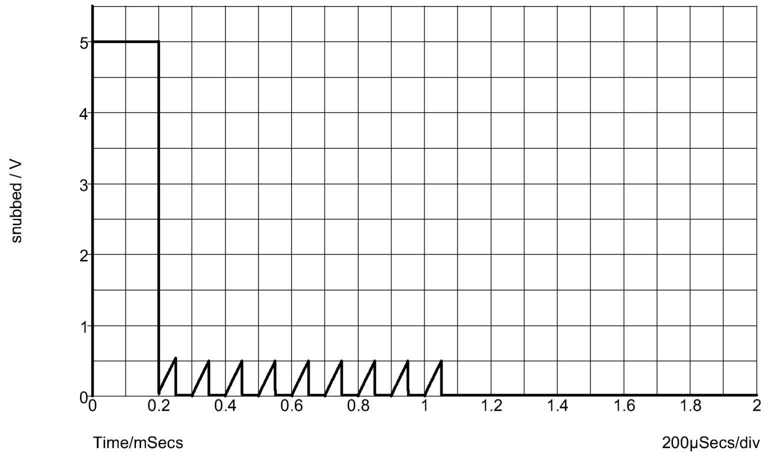
FIGURE 12.8D:



Re-arranging the parts gives a speed improvement. This scheme discharges the capacitor rapidly, limited by R1 and the switch resistance. It is therefore not a good idea to omit R1 when using a low impedance switch; the current in the switch and the capacitor are no longer defined, and damage or RFI may result. Carbon loaded membrane switches with >10 Ω ON-resistance will be OK without R1.

R1 should be low enough to discharge the capacitor on the first switch bounce. The inter-bounce recovery voltage must be made lower than the gate hysteresis.

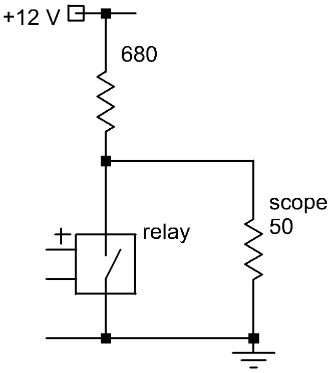
FIGURE 12.8E:



The recovery pulses are a bit large in this simulation, as they are intended to show the nature of the circuit. The key thing to note is that the logic level changes immediately and does not require the switch to have stopped bouncing. This circuit switches a higher current than the previous one and so will generate slightly worse RF interference, especially if tracked badly on a PCB.

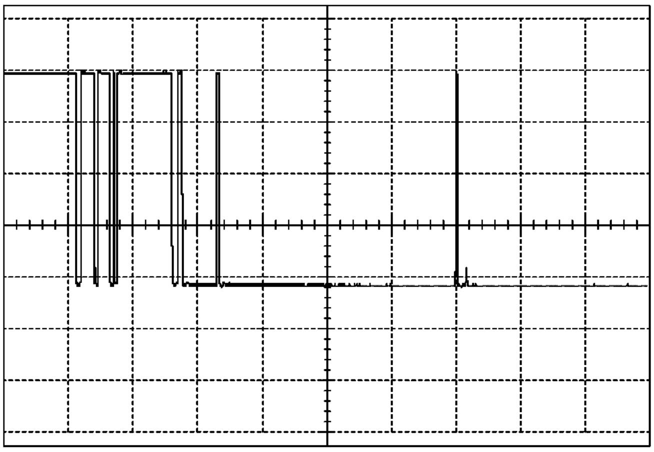
The loop formed by the switch and the snubber [C1 and R1] must have a small area, minimising the magnetic loop area. If the ends of the switch and the capacitor are carelessly connected to a ground plane or a ground track, voltage spikes will be injected in to that line.

FIGURE 12.8F:



Here is an actual contact bounce test circuit used on a Matsushita TQ2-12V relay running at 12 V. It was measured by shorting the 50 Ω input of a scope, which was otherwise pulled up to +12 V through a 680 Ω resistor.

FIGURE 12.8G:

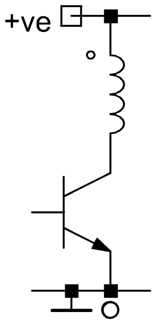


Measured result of bounce test:
200 mV/div.
50 µs/div.

Notice that it takes quite some time for the contact to settle down to a stable ON-state. It should also be clear that the bounce characteristic is far from regular.

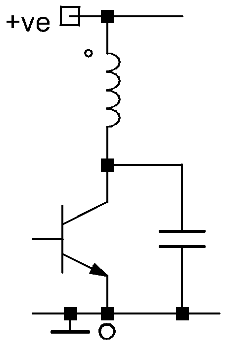
12.9 Coil Drive Circuits

FIGURE 12.9A:



A relay coil is most appropriately drawn as an inductor; you then can't forget its inductive nature. When the transistor turns off, the current in the relay coil will cause the voltage across the transistor to rise very rapidly to some value well above the power rail. In practice it will probably rise sufficiently to cause the transistor to exceed its breakdown rating. This is not necessarily fatal for the transistor; it depends on the type of transistor and the type of relay. Relying on these unspecified parameters of the devices is poor practice and is not worth the risk.

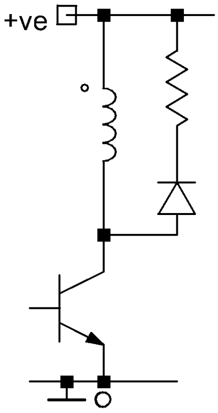
FIGURE 12.9B:



Here the inductive 'flyback pulse' is empirically {experimentally} tuned with the capacitor to not exceed the breakdown voltage of either the capacitor or the transistor. It gives a fast turn-off of the relay and additionally decouples the coil for interference suppression purposes. The capacitor value, found by experiment for the particular relay type, is likely to be 100 nF or greater.

The inductor current changes according to $V = L \cdot \frac{dI}{dt}$. Lots of reverse voltage across the coil makes the current drop to zero quickly.

FIGURE 12.9C:



This is the 'text-book' technique, but usually the resistor is left out. The diode has a reverse voltage rating at least equal to the power supply. Its forward current rating would probably exceed the relay coil current, but if not, ensure the diode can withstand the transient current magnitude and duration.

Estimate the resistor value by deciding how high a peak voltage you want on the transistor and then subtract the power rail voltage. Divide this resultant voltage by the maximum relay coil current and you get the maximum resistor value.

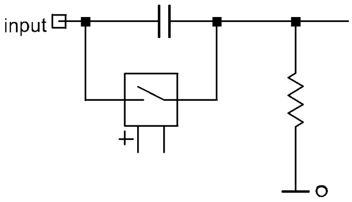
The resistor is only used if you need the relay to switch OFF quickly. If, instead, you want the relay to switch ON faster than normal, apply more voltage.

In the world of industrial ink-jet printers, where ink is dispensed by a solenoid valve, turning the solenoid on and off at high speed (5 kHz) is a priority. The solenoid is connected across an unreasonably high voltage rail, but only for a brief period, allowing the current to ramp up to the magnetic saturation point as quickly as possible. The drive is then chopped (PWM) to prevent the current continuing to rise. The flyback diode is connected to an even higher voltage power rail, or to a zener diode, to dissipate the stored energy as rapidly as possible.

12.10 AC/DC Coupling

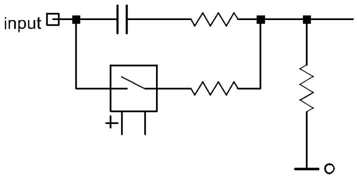
Switching between AC and DC coupling is often required in instrumentation & measurement applications. The AC coupling is obviously done with a capacitor and it is often convenient to use a relay to switch to DC coupling since the capacitor may be operating at hundreds of volts. This is a relatively simple function and yet it is easy to do wrong.

FIGURE 12.10A:



***EX 12.10.1:** What is wrong with this simple AC/DC coupling circuit, especially when used at several hundred volts?

FIGURE 12.10B:

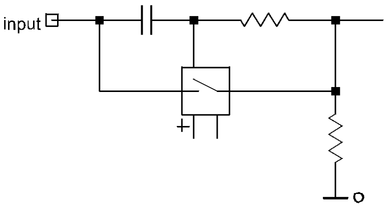


***EX 12.10.2:** Other than the obvious increased power/voltage handling capability, why use two resistors to protect the relay contacts?

The best way to do the AC/DC coupling switching is by the use of a changeover relay contact. Even then people still wire them up incorrectly! The two key mistakes are:

1) shorting out the capacitor directly, thereby burning out the relay contact and possibly damaging the capacitor.
2) using the relay as a MUX, selecting either the AC or DC path, but then leaving the capacitor charged so that when it gets reconnected a surge current out of the equipment becomes possible.

FIGURE 12.10C:



This is the optimum true AC/DC coupling circuit. There is minimal AC/DC coupling error because no extra resistance has been inserted in the AC or DC paths. The relay does not short-out the capacitor and the capacitor is discharged at a leisurely rate, ensuring that the charge on the capacitor is gently dissipated.

CH13: rating and de-rating

13.1 Introduction to Derating

A component is given a rating by a manufacturer. This may relate to power, voltage, current, tensile stress &c, and somewhere in there you will also find temperature. It is your job to select the **appropriate** rating for the component to use in your design.

In the term "de-rating" (the hyphen being optional) the *de-* prefix means *down* or *lower*. Having been given a manufacturer's spec under a certain set of conditions, you may decide to *derate* the component for longer life and therefore greater reliability. There is another use of this term though, and that is when the manufacturer specifies the device at a certain operating ambient temperature, but you wish to run it at a higher ambient temperature. Less power dissipation is allowed at this elevated temperature. The fact that both of these concepts are called derating can be confusing.

I am going to present some derating figures for reliable operation. You are free to argue with these numbers; I have even left a column for you to write in your own figures as your own experience leads you to a better answer. The point of the exercise is to realise consciously what it is you are doing with the derating and to apply it in a consistent manner.

These figures are for commercial designs and are not suitable for life support or safety critical applications. These figures are for a cost effective solution, rather than with ultimate reliability in mind. You will in any case need specialist training if you intend to be dealing with life support applications.

13.2 Safety Ratings

At one time a factor of two derating was used as a criterion to suggest that a component would be 'unlikely to fail' as far as a safety analysis was concerned. A resistor that was run at half of its power rating could be deemed 'unlikely to fail'; open-circuit and short-circuit tests would not need to be done on such a derated component. This procedure is no longer considered satisfactory.

The current procedure is as follows. Look over the circuit diagram and consider every component, one at a time, to see what might happen if that component were made open-circuit or short-circuit. This is known as a *single-fault* condition. If under this single-fault condition the circuit becomes unsafe, changes have to be made.

In any system, there are bound to be components directly across internal power rails. These components would include decoupling capacitors for example. You must investigate what might happen if a device were left open-circuit, or if it were to become short-circuit, initially as a theoretical exercise. If there is any slight chance of a problem you perform a test.

Applying a short-circuit across the power rail, you might find that everything is ok for the first few minutes, then the bridge rectifier overheats and goes short-circuit. Now this is all part of the single-fault condition. The bridge rectifier has failed *as a consequence* of the earlier failure.

The bridge rectifier is now shorting-out part of the secondary winding of the main power transformer. Unfortunately the transformer is rated for very high power, so the primary fuse is large and doesn't blow. The transformer is left overheating for a further few hours. Yes hours; transformers have very long time-constants, and do need to be tested for hours. The transformer insulation now breaks down and allows the primary winding to come into contact with the secondary winding. This short within the transformer will make the previously isolated secondary supply live. It is not guaranteed that the primary fuse will blow, but if anybody were to be in contact with any circuitry supplied by this transformer, the risk of electric shock would be high.

There are a few particular things you should be looking for :

☹ electrical hazards, where >42 V peak can connect to a person.
☹ smoke and poisonous fumes.
☹ exploding debris which could injury somebody.
☹ loud scary bangs (you might drop something if scared by a loud bang)
☹ accessible hot objects.
☹ nasty pointy objects that scrape or stab people.
☹ heavy or powerful objects which nip or crush body parts.

You need to guarantee that these things cannot occur under single-fault conditions, saying so in writing, *and* being held liable for them if they are subsequently found to occur.

Now you are thinking that my previous scenario was a very unlikely situation, and you are right for this day and age. However, before the safety standards were developed, it was not at all unusual for people to get shocks off of electrical equipment. That was why the standards were developed in the first place. The safety standards encapsulate 'best practice' and it is not only good sense to follow them, it is also a *legal requirement.*

Let me give you a quick overview of how that previous nasty scenario could have been avoided. In what follows, *earth* means the *circuit protective conductor* of the incoming AC mains supply.

☺ Earth the user control panel so that fault currents are shunted to earth directly and not via the user. The earth path would ordinarily be able to take 25 A for 1 minute and have an impedance of less than 100 mΩ.
☺ Put a correctly rated fuse in the AC ***mains*** supply line.
☺ Use an earthed inter-winding screen between the primary and secondary windings of the transformer. In case of a short-circuit, the fault current is shunted to earth.
☺ Imbed a thermal cut-out within the transformer insulation. Wire it in series with the transformer primary so that if the transformer overheats, the AC power is shut off.
☺ Earth one side of the regulated supply fed from the transformer. This would ordinarily be used as the system 0 V. Any short to the live supply would be shunted to earth.

That is not a comprehensive list of fixes for the problem and you might do several of those actions rather than just one. I just want you to get the idea of what can go wrong and how you have to *think* in order to not have a problem when it comes to the point of having a safety check done.

Now you should be able to see that using an electrolytic capacitor with double the necessary working ripple current rating and double the operating voltage rating is not a guarantee that the component will not fail. Better guarantees are needed. The rules currently state that you do not need to test 'high integrity components' on this open and short-circuit test. What constitutes a 'high integrity component' is an area which will be developed over time. There are specific rules for transformers and some other components, but you will need to check on what the latest rules are as you do each new design.

Now I have suggested that you only need to do the open-circuit and short-circuit tests where there might be a problem. The actual rule is that you must test *unless* you can prove that there would be no problem if the test were to be done. For a PCB assembly with lots of low power R's, L's, C's, diodes, transistors, ICs &c, you can usually just look at the board and say there is no problem. You would however open-circuit and short-circuit the power supplies to the board to see if anything else catches fire or explodes.

You should take particular interest in the high power areas, the high voltage areas and the earth {ground; protective conductor} path. These are the biggest sources of possible problems.

Sometimes it is difficult to imagine a particular fault condition occurring, but it is easier to just add an extra component than it is to worry about it. Nickel metal hydride batteries, or any rechargeable battery type for that matter, are a nice target for safety standards. If the battery gets overcharged, will it explode? Well, a little button cell is unlikely to explode and emit toxic amounts of fumes when it is being charged from a 5 V rail through a 2K2 resistor, but if you think about what might happen if the resistor should become short-circuited, it would be difficult to *guarantee* that nothing unpleasant could happen. In a situation like this it is easier to just put two resistors in series. If one fails short-circuit then there is still no problem. You have increased the equipment cost only marginally, but you have saved yourself a significant amount of testing. I would recommend this type of solution to you. It is arguably a bit silly, but there is no denying that it meets the standard.

If you work in the Nuclear industry, aerospace industry, or on life support equipment, then a single-fault analysis is clearly far from adequate. You are always free to consider dual fault situations, if you wish, to add extra safety margin.

13.3 More Derating

Let's talk about the non-safety related stuff. You want the circuit to be adequately reliable and that includes looking at the worst case conditions. All derating of components has to consider what the worst case condition is, and for how long the unit will be subjected to these conditions. If I were considering a user operated switch, I would look at it as if the user were sitting there using the switch for 8 hours a day, 5 days week, 52 weeks of the year, for five years. If it can stand that sort of operation it is certainly adequate. Arguably it is over-designed but that comes down to a cost-reliability trade-off. I prefer the user to get worn out before the switch, and I don't like doing repairs under warrantee.

If the result of your calculations proves expensive to achieve, you do what any good engineer does, you make an engineering compromise taking the cost into account. Some edge connector sockets on PC systems, for example, can only take 5 insertions. That is a very small number when there are problems with the assembly and it needs to be pulled in and out a few times.

Each component has its own rules for derating and I will discuss each one individually before compiling a table of numerical values.

In the discussions that follow, the deratings are from the temperature derated spec. Let me explain that fully with an exercise.

EX 13.3.1: A resistor is rated at 250 mW in a 70°C ambient. The manufacturer tells you to derate this linearly to zero at 150°C.

a) What is the manufacturer's derated spec in a 100°C local ambient?
b) The derating figure given in the table is 80% for continuous use. What is the maximum power you should dissipate in it given this 100°C local ambient?

Resistors are generally not affected by voltage too heavily. If you put a bit too much voltage on them, then their resistance may change a little, but it is more of a gradual change. They are often rated for overload at double their maximum rating anyway, so I do not worry about their voltage rating unduly. When suppressing electrostatic discharge, resistors get overloaded horribly according to their specs, but in practice they are able to take it.

Having said that, you will find that resistors that are used in mains {off-line} applications are very prone to failure if they are connected directly across the mains or nearly across the mains. For 230 V operation you might reasonably think that $230\times\sqrt{2}\times1.2=390\,\text{V}$ was perfectly adequate. You would not only be wrong, the cost of this wrongness could be the downfall of your employer and/or of your career. The peak voltage across the mains is much higher than the peak of the RMS voltage would suggest. Check back on the section on Class X capacitors for more details. If you want an idea of how much voltage a component is expected to withstand in a mains circuit then look at the safety standards such as IEC61010-1 and IEC60950. The ***basic insulation*** test requirements are a good guide to the expected surge voltages on the mains. When I say "peak voltage" in the derating table, I really do mean peak.

Now you might reasonably ask why I should bother coming up with a table of additional derating values. After all, the manufacturer has given them ratings and told you how to derate for ambient. Why should you add any more? That is a very good question, and if you hadn't thought of it then you should have. You must challenge these ideas rather than just accepting them obediently. This additional derating is potentially costing money.

It is a difficult area. When dealing with commodity parts like resistors and ceramic capacitors you may find that some manufacturers are more diligent than others. Whilst some will have plenty of margin in hand for the figures they give, others might not be so generous. Less diligent manufacturers may well charge less for their parts and component buyers, wishing to save money, will buy the cheapest. It is therefore not safe to assume that the spec has anything "in hand" {excess performance over and above the spec}.

The other major consideration is temperature. This is not an exact science and don't let anyone fool you into thinking that it is. You have a multitude of components on a circuit board. Their proximity to each other is not something that you specify on the circuit diagram. Therefore all the hot components can end up in one area of the board. It is remarkably difficult to say what the ‘local ambient’ is for any individual component. Initially you just have to ‘guess’ and use a bit of judgement. The other thing that you can and should do is to run your hands over the working prototype and see which bits get hot. [***SAFETY WARNING:*** Ensure that you only do this on boards with less than 40 V on them, remembering to include voltages generated locally by switched-mode supplies.]

Ok, the correct thing to do is to use a thermal imaging camera, but these are currently too expensive to buy (unless you work for a very large company). Maybe you can rent one; they do give excellent results. A cheaper alternative is an infra-red spot-reading device. This focuses on a small region and can therefore sense the temperature of an individual resistor, for example. The only slight drawback is ***emissivity*** calibration. You need to correct for the emissivity of the surface of the measured device. However, this is easily accomplished by uniformly heating the whole board up to a known temperature in an oven. This allows the emissivity of the component to be corrected for. [Just change the emissivity correction factor until the infra-red device reads correctly at the elevated temperature.]

Those parts that are running hot can be probed with a bare thermocouple (a thermocouple probe is often too big and bulky to get an accurate reading on a small device). If the temperature is getting close to a limit then you will need to bond the thermocouple to the device to get a better thermal connection. You can also get special paint which changes colour at a specific temperature. This paint is ideal where the component is in a circuit which has fast moving, high voltage signals on it, making accurate thermocouple measurements both difficult and dangerous.

A good reason for using a larger amount of derating is uncertainty. If you cannot be sure how much power is going into a particular device that is (electrically) swinging up and down all over the place, then it is better to “play safe” than overrun the component.

What about fan cooling? For small devices on a board I do not like to rely on fan cooling to save them from destruction. I tend to use the fan cooling to improve their reliability. Again it is very difficult to calculate, or even estimate, what air flow you are actually going to get at board level and what effect this will have on the components. You can use an air flow probe on the completed product, but by then it can be a bit late to be changing the component types. And for transportable equipment you have the additional problem that the equipment may be used in an infinite number of different orientations. You may write in your user manual that the equipment has to be used laying flat on the bench, but *homo illiterus* [†] may not bother to read that part. The calibration may be out of spec, but the instrument should not overheat and blow up {fail}.

What about air pressure? Did you realise that both forced cooling and naturally convected cooling require air, and that at high altitudes there is less of it around? More specifically, it is a good approximation to say that atmospheric pressure drops 100 mbar/km up to 5 km. In other words, at 5000 m the atmospheric pressure is halved. For most land based equipment, the maximum operating altitude is likely to be around the 2000 m mark. Equipment operating in these high altitude environments will overheat if it is running too close to its thermal limit. At 2000 m, the reduction in cooling for small pieces of equipment (<1 m per side) is likely to be in the range of 10-20%. This is why you see environmental ratings for equipment which refer to operating altitudes.

[†] Latin sounding name, suggesting a person who can’t read.

Whilst aircraft cabins are not allowed to go below a pressure equivalent to an altitude of 2400 m, you have to consider if the equipment needs to function in emergency decompression situations. If the equipment is for use by an operator, as opposed to remote control, then it is useful to know that a person used to living at sea level would suffer from hypoxia {oxygen deprivation} at an altitude of >3000 m. Non-operating altitude specs account for the possibility of transportation in unpressurised aircraft holds.

For heavy duty components with heatsinks, or expensive components with heatsinks, fit an over-temperature cut-out of some description. These are very easy because the temperature difference between the fan running and not running will most likely be at least 20°C. Not only can the fan fail, but the filter can become blocked, or *homo stupidus* may place the fan inlet right next to a wall, thereby preventing adequate airflow.

13.4 Derating Table

This table is to give you ideas rather than to constrain you. It should not be used as a rigid code to be followed on pain of dismissal. Get a feeling for the subject and make compromises where necessary. On one switched-mode design the peak switch voltage was specified as 35 V. I ran it up to 33 V repetitive peak transients, but 100% tested to make sure that they did not go above 33 V. Not nice, but it was a low volume unit, and clamping the transient harder would not have been either easy or efficient.

In some countries, and for some applications, safety checks by nationally recognised laboratories will be mandatory {compulsory}. You must check your product to see if it falls into one of these mandatory categories.

COMPONENT	sub-type	spec	SEEKrets derating	your derating
resistor		peak voltage	90%	
	0.25 W	peak voltage – static discharge (However, a resistance change of 10% must not cause a problem in the circuit.)	±15 kV	
	0805 size	peak voltage – static discharge	±8 kV	
		power continuous	80%	
		power pulsed	100%	
		power repetitive peaks	use the formula	
capacitor	electrolytic	ripple current	70%	
	electrolytic	peak voltage	95%	
	others	peak voltage	80%	
		dV/dt	80%	
		mains -across line	Class X	
		Mains - to ground	Class Y	
diode	ultra fast	current	body temperature of 80%	
	ultra fast	mean current	80%	
	mains rectifier 230 V	voltage	≥1000 V rating required	
	zener	power	80%	
	0.25 W zener	Static discharge	±15 kV	
bipolar transistor		reverse b-e when used as an amplifier	|spec| −0.5 V	
		reverse b-e when used as an LF switch	<10 mW	
MOSFET		voltage d-s	80%	
		power	body temp of 80%	
mains switch		use an approved part with lots of international safety approval stamps		
Mains wiring		use “Tri-rated” wire.(UL / CSA / IEC)		
mains transformer		specified by safety standards		
mains inlet		use an approved part with lots of international safety approval stamps		
fan		must not burn out even when rotor deliberately jammed		

CH14: circuit principles

14.1 Introduction

Problems in text books can be mathematically complex, but are straightforward in the sense that they relate to a specific chapter and topic in the book. In real life, difficulties arise because there is nobody around to open the appropriate text book at the right page, or even the right chapter. It is *much* easier when you know what type of solution to apply.

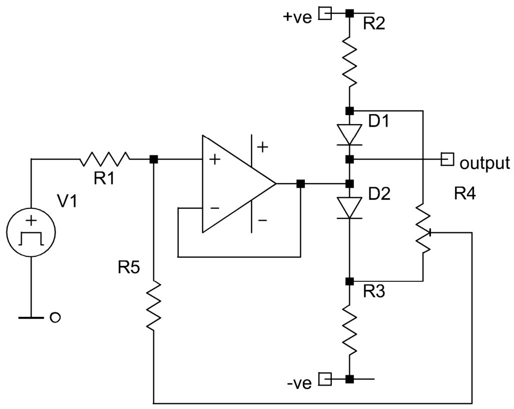
***EX 14.1.1:** You are designing a power system for a cryogenic quantum demodulator.[†] Your system requires at least three separate switched-mode power supplies. Your boss has asked your opinion about this in terms of noise from the supplies. He wants to know if the power supply switching frequencies should be synchronised or left to free run. Before you can answer this, however, a colleague "helps you out" as he is passing on the way to the coffee machine. He just shouts out that uncorrelated sources are summed in an RSS fashion, whereas correlated sources add. Your boss is not one to be swayed by random gossip, and since you are the engineer in charge, he wants your opinion. Fortunately you have a bit of time to think about it, as he has to go to a meeting. What is your professional advice?

14.2 The Bootstrap

A bootstrap is a leather loop attached to the back of a boot which can be pulled to help get the boot onto the foot. It would be ridiculous to think of somebody pulling on their own bootstraps and thereby lifting themselves off the ground, but nevertheless you may hear phrases like "pull yourself up by your own bootstraps", meaning to improve your situation in life by your own efforts.

This "amusing" concept has been carried into the electronics world. You can't call it a circuit, or sub-circuit; it is the idea that is the common factor, not the circuit. The technique seems to have been developed during the war years of 1939-1945, but is not credited to a specific designer.

FIGURE 14.2A:



V1 and R1 represent a source with a significant output resistance. The amplifier is a *unity gain follower*. If the amplifier bias current is too high you can reduce its effect by means of a *bootstrapped* bias current cancellation scheme. Regardless of the input voltage, R4 is always at the same voltage relative to the opamp input. Bias current can therefore be injected via R5, where R5 >> R1.

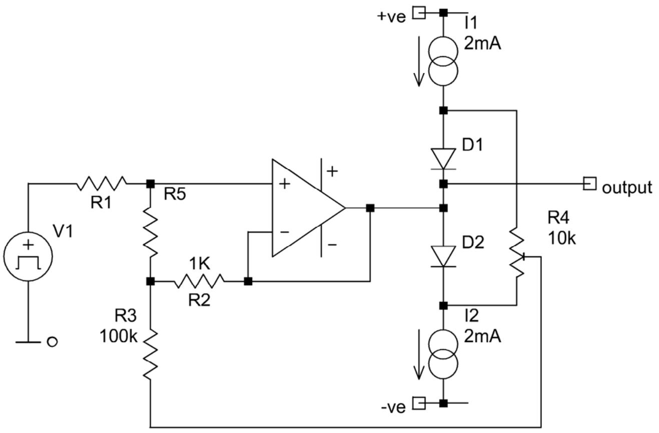
[†] Invented device type.

EX 14.2.1: R1 is 1000 kΩ. The amplifier has a bias current of ±3000 pA. Neglect drift of the bias current with time or temperature, and neglect any offset or CMRR error in the amplifier:

a) What is the maximum offset caused by the (uncompensated) bias current?
b) Select a suitable value for R5.

This scheme is not suitable for all applications because R5 ends up considerably larger than R1. The solution is to make a divider with respect to the amplifier's inverting input.

FIGURE 14.2B:



This scheme allows R5 to be 100× smaller in resistance, an important consideration when its value might otherwise exceed hundreds of megohms.

***EX 14.2.2:**

a) What difference does the input offset voltage of the amplifier make to this circuit?
b) What difference does the input offset TC of the amplifier make?
c) What difference does the CMRR of the amplifier make?
d) The input offset of the amplifier is ≤ ±1 mV and its bias current is ≤ ±500 pA. What is an appropriate value for R5?

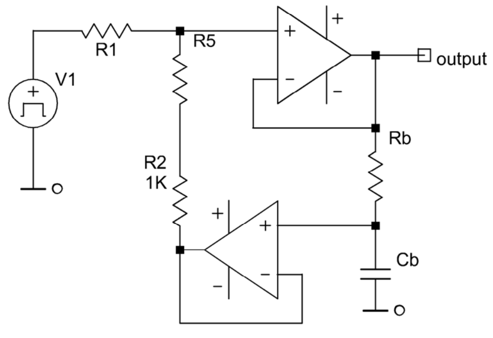
The answer to that question should have shown you that there are several possible problems with the circuit as drawn. The division ratio of 101:1 is a bit high for the pot adjustment range of ±700 mV. If the division ratio is reduced, more span is available, but it also means that the bias current resistor needs to be larger. The alternative approach of using reverse-biassed zeners instead of forward biassed diodes gives a more stable adjustment scheme, at the expense of limiting the output swing of the amplifier. Due to the relatively high dynamic impedance of zeners, it is possible to obtain better results, albeit at increased cost, using 3-terminal voltage regulators, low impedance shunt regulator ICs or reference voltage generators.

In practice, the feedback division ratio needs to be in the 10 to 100 range, and the voltage sources need to be between 700 mV and 5 V when using ±12 V or ±15 V rails. Start from the value of R5. Use the highest value you can cheaply obtain and work everything else out from that starting point.

Ideally the bootstrap would be much faster than the signal, giving minimal phase shift over the frequency range being used. The question you should be asking is just how fast does the amplifier have to be in order to not have an adverse effect on the circuit's performance. If the bootstrap is 10× the bandwidth of the input signal then you might *assume* that this was acceptable.

To make this analysis I am going to simplify the problem by asserting that the loss of bandwidth occurs in the amplifier and not in the level shifting circuitry. This should be a reasonable assertion for the simple network shown. I am going to further simplify the problem by looking at the finite bandwidth of the amplifier as being a simple single-pole roll-off. Again this should be reasonable because I am not going to be looking at the response past the 3 dB corner.

FIGURE 14.2C:



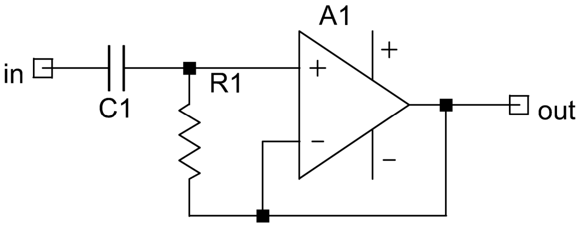
Rb and Cb model the band-limiting effect of the buffer amplifier. R2 has been shown only for the purpose of orientation to the circuit. In normal use the value of R2 should be at least 100× lower than R5 making it unimportant in an analysis.

***EX 14.2.3:**

a) Correct the error in the above circuit model. Hint: *what about the output*?
b) Having corrected the above model, develop a transfer function for the circuit using the symbol B for the 3 dB bandwidth of the Rb–Cb network. Neglect R2 and any errors in the buffer amplifiers.

The bootstrap technique is often used to increase the input impedance of an AC coupled amplifier. In general the active device needs a bias network and the input coupling capacitor 'sees' this bias network in parallel with the input impedance of the active device. This bias network will therefore be the dominant input load for FET based amplifiers below a few hundred kilohertz.

FIGURE 14.2D:



In this simplified equivalent circuit, the voltage gain of A1 is slightly less than unity. Ordinarily the bias network would be a potential divider from the power rails and the bootstrap would couple into it.

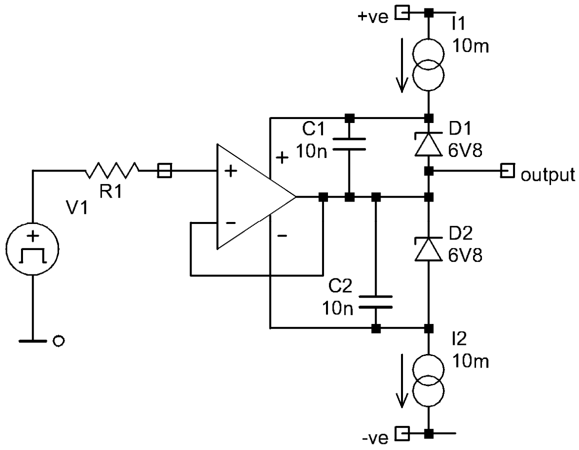
This simple equivalent gives the relevant characteristics of *all* bootstraps of this type.

***EX 14.2.4**: Write the voltage gain of A1 as $(1-\delta)$, where δ is small ($\delta < 0.1$) and positive. Neglect any phase shift or finite bandwidth in A1.

a) What is the effective value of R due to the bootstrap?
b) What happens to A1's output voltage noise as a function of frequency?

If you recall the section on opamps, you will remember that in voltage-follower mode the key error is due to the finite common mode rejection ratio of the opamp. The load resistance also has an effect due to the opamp's finite gain at the signal frequency.

FIGURE 14.2E:



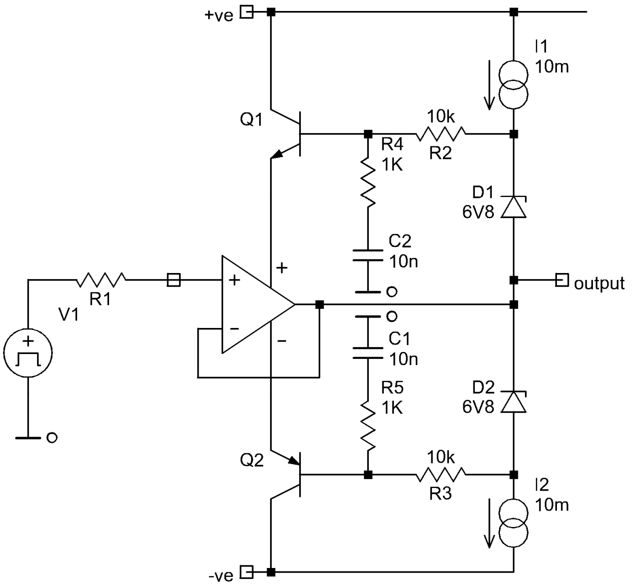
By bootstrapping the power rails, the inputs are not changing with respect to the power rails. The finite common-mode rejection ratio does not causing an error when the input signal changes.

Notice that the output is not changing with respect to the power rails so the gain of the amplifier is not causing an error.

This circuit does not work.

***EX 14.2.5:** Why doesn't it work?

FIGURE 14.2F:

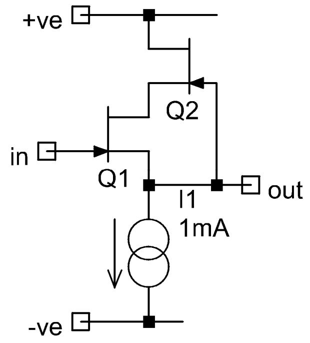


This next scheme is workable because the current from the power supply is not all coming through a single constant-current device. Furthermore the phase shift around the loop is controlled by RC networks to prevent oscillation. The component values need to be tuned {adjusted} for optimal response. A fast edge (eg 1 μs risetime) should be applied to the input and the values adjusted until the output ***ringing*** is minimised.

You want the fastest bootstrap loop possible, without excessive overshoot or ringing. The accuracy of this circuit will be very much better than the basic amplifier [say 10× better], but a lot of circuit complexity has been added. You would only do this if you really needed the extra accuracy or better repeatability with device variations. What you get is a much larger and better defined CMRR. Whilst the CMRR of opamps is generally a *chord slope* value over a large range, this averaged value does not show you the non-linear behaviour of the CMRR with input voltage.

Once you have understood the circuit principle, you can adapt it your application. For example, using JFETs, a simple self-biasing *source-follower* bootstrap is easy to make.

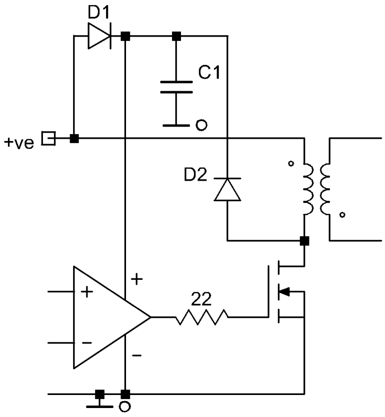
FIGURE 14.2G:



This is a very neat scheme because the bootstrap transistor, Q2, needs no bias components at all. An N-channel JFET requires the gate to be more negative than the source in order to turn the channel off. Run Q2 at considerably less than its I_{DSS} in order to get at least a volt or so across the drain-source of Q1. Note that Q1 could be an NPN bipolar transistor and the bootstrap would work just as well.

Another form of bootstrap is for a power supply. On switched-mode supplies it is easy to tap off of the switching circuit in order to get a higher gate drive voltage; this makes the MOSFET more efficient.

FIGURE 14.2H:



The +ve supply might be 3 V and the MOSFET would not be running very efficiently. It must still be able to run at this low voltage, however, or the circuit won't be able to start. Initially the supply comes from D1. When the circuit starts switching there is additional voltage available through D2 and the power rail of the switching device can be brought up to perhaps 10 V or more. This would allow the MOSFET to be turned-on harder and therefore to be more efficient.

14.3 Guarding & Cross-Talk

Guarding means different things to different people. For an *ATE* [automatic test equipment] technician, guarding means the action of measuring a component "in circuit", meaning that there are other components wired in a complex network across the component under test. The guarding to be discussed in this section is for less definite leakage paths.

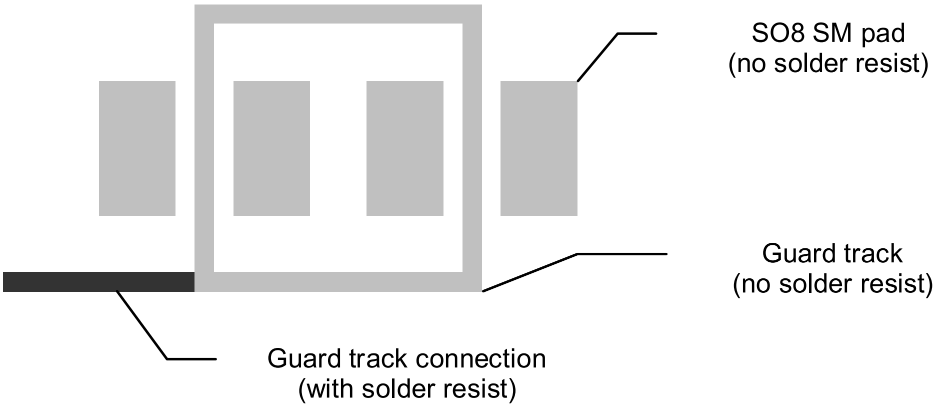
Guarding is the process by which stray leakage currents are diverted from critical *nodes* in a circuit. It is a very simple, but powerful technique. Once you have mastered the basics, it will seem so obvious you will wonder why anybody had to write it down! This is why I like giving you problems to solve *before* telling you how to solve them. It makes you realise that there *is* something to know.

Guarding is necessary when measuring or routing *low level* currents, for example those with resolution required below 100 pA. This applies to circuits in physically dry conditions {*not moist or wet*} without excess dirt/dust present. If you have PCBs that may be covered in dirt, dust and condensation, some sort of guarding may be even more important that usual.

For a printed circuit board, a *guard track*, connected to a suitable voltage, is routed around the sensitive circuit. The key point is to have the solder resist removed from this guard track. You are trying to stop leakage across the contaminated surface of the PCB, not the inside of the solder resist! If the contamination is heavy, you will need a wider guard track. Think of a guard track in the region of 0.006 inch (0.15 mm) to 0.200 inch (5 mm).

Guarding is particularly difficult around surface mount ICs because the guard track can short to the adjacent pins if excess solder is used. Nevertheless, the guard track must **not** have solder resist over it. Such guarding is problematic on standard SO8 SM packages and impractical on any finer pitch device.

FIGURE 14.3A:



The pads within the guarded ring are protected against leakage currents from adjacent pads and tracks. It is much easier to guard DIP packages and you may need to revert to wire-ended components at a few critical points in the circuit in order to get a design which is manufacturable. In any case, DIP parts sometimes have a tighter spec in terms of offset voltage, for example.

Another possibility is to mill out a couple of slots between the surface mount pads, rather than putting a guard track between them. If the PCB manufacturer is able to mill out these fine slots, there will be less production difficulties.

It may be that guarding is not going to solve your problems because the condensation is too great, or the layer of dirt is too great. In this case you will need to either coat the board with a thin insulating *conformal coating* or encapsulate the board in silicone rubber followed by epoxy resin [known as *potting the circuit*].

Potting and conformal coating both have the major drawback of making it very difficult to service the unit, but they are the only practical solutions for extreme environments (other than putting the circuit in its own gasket-sealed box).

If the circuit cannot be guarded because there is no suitable guard voltage, and none can be obtained, then in addition to the technique of milling out sections of the PCB, another solution is to stand the critical circuitry up off of the circuit board using PTFE insulated solder cups or pins. Obviously this technique requires wire-ended components and incurs increased hand labour costs.

To get you started, assume that the guard track is half way between the leaking area and the area being protected. Assume a resistance for the leakage path and you have your equivalent circuit. The leakage path should be >1 GΩ for a clean PCB, but it could drop to <1 MΩ in the presence of moisture and dirt. It is a question of how much exposed conductor is available to source the initial leakage current.

EX 14.3.1:

A guard track is run around the inverting input node (junction of R1 & R2) stopping leakage across the board affecting the gain and offset accuracy.

Where should the track be connected to?

FIGURE 14.3B:

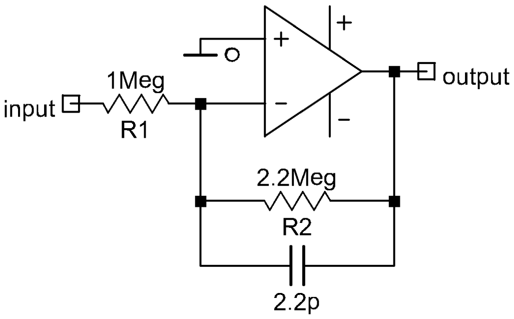
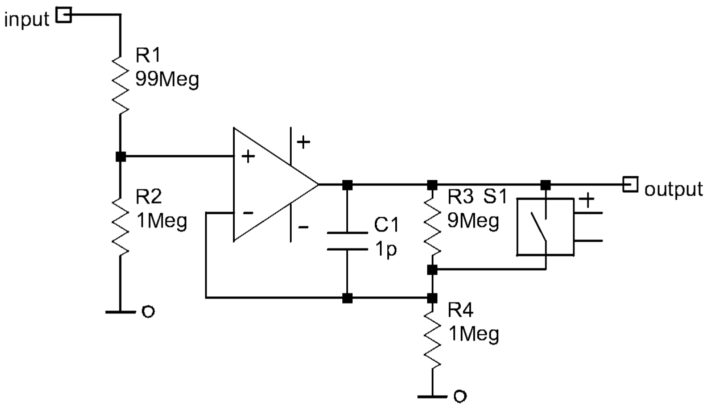


FIGURE 14.3C:



***EX 14.3.2**: This is a switched gain amplifier with very high input impedance. Unfortunately the leakage path across R1 is found to be up to 10 MΩ due to environmental considerations. Your boss suggests you guard the leakage path in the input attenuator. What is your response?

By now you should have realised that the bootstrap and the guard are intimately linked. A guard is bootstrapped in many circumstances.

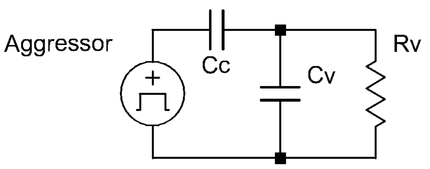
Cross-talk is closely related to the subject of guarding and also to the subject of shielding, covered in the next section. Whilst cross-talk can occur because of common-impedance coupling in ground or power tracks, the rest of this section is devoted to cross-coupling by some electric/magnetic/electro-magnetic mechanism, primarily between nearby conductors. Typically these conductors would be on a PCB, in a ribbon cable, or in a wiring loom {bundle}.

The easiest sort of cross-talk to understand and to eliminate is that due to capacitive coupling. This is sometimes called "electrostatic" coupling, although that is a bit of a misnomer since the signals are clearly changing! The term is intended to mean *quasi-static coupling* in which higher speed electromagnetic phenomena are not involved. Hence I shall stick with the name 'capacitive coupling' in order to avoid the cumbersome term electro-quasi-static coupling!

There are five key factors that need to be quantified in order to estimate the capacitive cross-talk effect between an 'aggressor' (source) and a 'victim' circuit.

1) The voltage change on the aggressor circuit.
2) The risetime of the voltage on the aggressor circuit.
3) The capacitance between the aggressor and victim circuits.
4) The effective input capacitance of the victim circuit.
5) The effective input resistance of the victim circuit.

FIGURE 14.3D:



The resulting equivalent circuit makes the mechanism much easier to see and understand. If there is a 10 V swing on the aggressor circuit, the maximum swing on the victim circuit is given by the division ratio of *C*c to *C*v. As an example, a 1 pF cross-coupling capacitor will not produce more than a 1 V swing in the victim circuit if *C*v is 9 pF.

However, if the risetime of the aggressor voltage is much slower than the time constant of the capacitive coupling, the peak pulse will not even reach the value calculated from the capacitor ratio. In this case the rising voltage gives a current, calculated from $I = C\dfrac{dV}{dt}$. Denoting the rise-time of the aggressor voltage as T_R, the cross-talk in the time domain is then given by the lesser of:

$$\boxed{V_{VICTIM} = \frac{\Delta V_{AGGRESSOR}}{1+\dfrac{Cv}{Cc}}} \quad \text{and} \quad \boxed{V_{VICTIM} \approx \Delta V_{AGGRESSOR} \times 2 \times \frac{CcRv}{T_R}}$$

A 10 V swing with a 1 µs risetime coupling via 1 pF into a 1K resistor gives 20 mV spikes; positive on the rising edge of the aggressor and negative on the falling edge.

For a long parallel run of tracks or wires, doubling the separation will not even halve *C*c, the capacitance dropping as a logarithmic function of distance. To get a large reduction in the cross-talk it is necessary to put a 'grounded' conductor between the victim and aggressor circuits, although this 'ground' could be a well-decoupled power or signal track.

It is noteworthy that the capacitively coupled crosstalk is necessarily of the same 'phase' as the aggressor signal. They both rise and fall at the same time. It is also noteworthy that the grounded guard track need only be connected at one end in order to be effective at reducing the cross-talk.

Capacitive cross-talk is very easy to understand and to remedy. Inductive cross-talk, on the other hand, is somewhat harder to understand. The inductive cross-talk mechanism is by mutual inductance between the aggressor and victim circuits.

Consider the case of two parallel wires in a ribbon cable. When a current flows in the aggressor circuit, a voltage is induced in the victim circuit, much in the same way as a voltage is induced in the secondary of a transformer when current flows in the primary. Using a low input impedance in the victim circuit has a minimal impact on the induced voltage.

A simple way to determine if the cross-talk mechanism is capacitive or inductive is to put a low value resistor across the far end of victim circuit, that is the end furthest from the receiving device. If the cross-talk is substantially reduced, the cross-talk mechanism is capacitive; if the cross-talk is unchanged or increased the mechanism is inductive. Another test is to look at the direction of the spikes. If the spikes are in-phase with the edges they are capacitive. If the spikes are in anti-phase they are inductively coupled.

Reducing inductive cross-talk consists of reducing the mutual inductance between the circuits, or slowing the rate of change of current in the aggressor circuit. Reducing the loop area in the aggressor and victim circuits, for example by using twisted pairs, will also be effective at reducing the mutual coupling.

Consider a ribbon cable 30 cm long with a wire pitch of 2 mm. Suppose a 100 mA load is switched with a risetime of 1 μs. The mutual inductance between the wires, using the formula from chapter 7, is found to be 280 nH. In rough figures the resulting induced voltage is $V = M\dfrac{di}{dt} \approx 280 \times 10^{-9} \times \dfrac{0.1}{1 \times 10^{-6}} = 28\text{ mV}$

Increasing the separation of long wires or traces only reduces the inductive cross-talk at a logarithmic rate. A guard track grounded at one end placed between the offending current and the victim circuit will have *no effect at all*. In order to reduce the magnetic coupling, induced current has to be allowed to flow in the guard track. This means that there has to be a low impedance wire loop present. You can think of this as putting a shorted-turn on a loosely coupled transformer.

A good way to keep the aggressor circuit separate from the other circuitry is to force the current to flow in a coaxial cable. Note that just connecting the coaxial screen at one end is entirely ineffective at preventing magnetic coupling. All the current must go down the inner and return down the outer in order to produce minimal external magnetic field.

When the risetime of the voltage (or current) edge is faster than the propagation delay down the cable (or track) these simple capacitive or inductive cross-talk mechanisms become inter-twined. The coupling mechanism is then electro-magnetic. The capacitance cannot be 'shorted out' because the edge does not occur at all points along the cable at the same time. Likewise the induction field cannot be shorted out.

The cross-talk is still improved by increased conductor spacing, grounded conductors between the aggressor and victim circuits, and lower impedance in the victim circuit, but if these remedies are insufficient, separate cables will be required.

Synchronous digital buses are fairly tolerant to cross-talk because they all change at once, settle, and then the clock changes state. Analog circuitry, on the other hand, is often inherently asynchronous; it does not have this luxury of being able to ignore cross-talk at particular time intervals. However, for a sampled analog system, if it can be arranged to sample the analog data during a 'quiet period', rather than during a noisy period, that type of scheme will give superior results. It is not unknown for a digital system to be temporarily shut-down for a brief period, allowing the analog signal to be sampled without undue digital interference.

14.4 Shielding

The shielding to be considered in this section is of three distinct types: magnetic field shielding, electric field shielding and RF shielding. Of these, magnetic field shielding is considerably more expensive and difficult than the others. Incidentally, the terms *shielding* and *screening* are synonymous, as are *shield* and *screen*.

Back to basics: Currents produce magnetic fields; voltages produce electric fields. Changing currents produce changing magnetic fields, which induce voltages into nearby conductors. Changing voltages produce changing electric fields, which capacitively couple into nearby conductors.

That is the elementary explanation. The previous paragraph missed out two other terms. Changing currents also produce changing electric fields. Changing voltages also produce changing magnetic fields.

Let me consolidate the two viewpoints: a changing current or a changing voltage will both produce an electromagnetic field. If the detector is sufficiently far away from the source, it will not be possible to tell whether the source is a changing current or a changing voltage.

Close to a ‘magnetic source’ (a changing current) the electromagnetic wave will have a very low ***wave impedance***, this being the ratio electric field intensity over magnetic field intensity. Close to an electric source the electromagnetic wave will have a very high wave impedance. Far away from either source, the wave impedance will be a constant 377 Ω.

Physicists sometimes call the close-in region the *Fresnel zone*; I prefer the term *near field*. Closer than $\lambda/2\pi$ is generally considered to be the near field; further out is the *far field*, the *Fraunhaufer zone*. (For antennas, use the ***Rayleigh Distance*** to define “near”.)

Frequency	Limit of Near Field
100 kHz	480 m
1 MHz	48 m
10 MHz	4.8 m
100 MHz	480 mm
1 GHz	48 mm
10 GHz	4.8 mm

When working in the near field, the nature of the source is critical to solving any particular shielding problem. For example, trying to shield the 50 kHz magnetic fields generated by switched-mode power supply components is not going to be successful unless highly conductive shielding is used, the source wave impedance being very low.

A 10 mΩ shield is not going to have a dramatic attenuation effect on a field whose wave impedance is also 10 mΩ. It can therefore be more effective to use high permeability magnetic material to divert magnetic fields around critical circuitry.

Lines of flux of any field will always tend to follow the easiest path. For magnetic flux this is a high permeability path; for electric flux it is a high electrical conductivity path; for heat flux it is a path of high thermal conductivity. Whatever the flux, shielding can be achieved by diverting the flux around the critical area through an “easier” path.

An important point to mention concerning high permeability magnetic paths is that if the magnetic field is above several tens of kilohertz, the electrical conductivity of the path is also critical. Iron is a high permeability material, but at 100 kHz a large mass of iron will *not* divert a magnetic field by being an easy path; the electrical conductivity is sufficiently high that the eddy currents generated will make the effective permeability very low. You will find that flux “would rather” go through air than through a large block of iron at around 100 kHz or so. Thus one could argue that the effective relative permeability of the bulk material is less than unity at this frequency.

You can test this for yourself very easily. Wind a coil of wire around a thin insulating former with about 1 cm internal diameter. Measure the coil’s impedance by measuring the voltage across it when fed from a 100 kHz generator via say a 1 kΩ resistor. Now insert a 1 cm diameter iron rod into the coil and see if the measured voltage goes up or down. At 100 Hz the voltage will go up because the inductance of the coil is enhanced by the iron. At 100 kHz, however, the voltage will go down because the iron restricts the amount of flux that can pass.

Using the ‘flux conduction’ idea, it should be evident that if a shield is discontinuous, problems will result. And this is in fact the case. A perfect conductor of flux is a perfect shield. Unfortunately it is inevitably necessary to open the shield and it is also usually necessary to have connections to the outside world. These are the weak points of the shield.

The easiest field to shield against is the electric field. Suppose a wire or PCB track has a changing voltage on it with a repetition rate below 1 MHz. If this occurs inside a small piece of equipment (<1 m on each side) you are guaranteed to be in the near field region, and the source impedance will be relatively high. The interference mechanism is by coupling of the electric field from the source to the victim circuitry; a stray capacitive coupling if you want to look at it in simple *lumped-element* terms. There are two ways of approaching this problem; you can look at the field and try to 'see' where it is going, or you can model the field as a capacitive coupling. You need to be able to use both methods.

In order to reduce the coupling you need to be able to visualise where the flux is going and to divert it. In the simplest case, if the victim node {wire; point; terminal} can see (by direct line of sight) the interference source, then the capacitive coupling is obvious. The effective capacitance can be ridiculously low (< 0.01 pF) and yet still cause a problem.

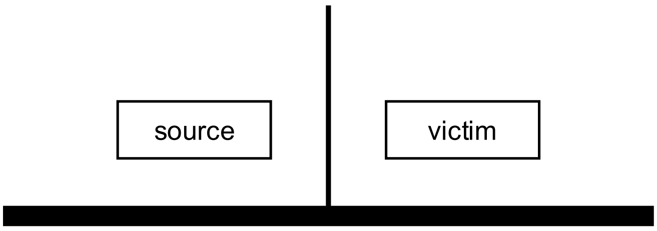
Consider a 10 V signal swing on the source node and have a node in the victim circuit with an impedance of 10 kΩ//10 pF. If there is a stray capacitance of 0.01 pF then the signal on the victim circuit will be 20 mV ptp. Each edge is differentiated, giving a +10 mV spike then a –10 mV spike. The source could have a long wire (>15 cm) or PCB track associated with it, and the victim circuit could also have some length. This means that the stray capacitance figure given could easily be exceeded by several orders of magnitude.

The first level of 'shielding' is to put a grounded track between the source and the victim; this is just the guard track from the previous section, but now there is no need to remove the solder resist from the track. This simple track may be all the 'shielding' that is necessary. You could say that the crosstalk has been reduced. There is no need to build huge screened boxes if a simple guard track will do the job.

Sometimes the problem can be overcome just by increasing the track or component spacing. This is particularly true when the source and victim are more like point sources rather than long wires. The rate of decrease of the field with distance will then be faster than the inverse log law found with parallel wires.

Air, or more particularly distance, can be one of the simplest "shielding" methods to employ. It is when this first 'opening shot' fails that you move on to the simple conducting barrier screen. You have now moved to a three dimensional model.

FIGURE 14.4A:



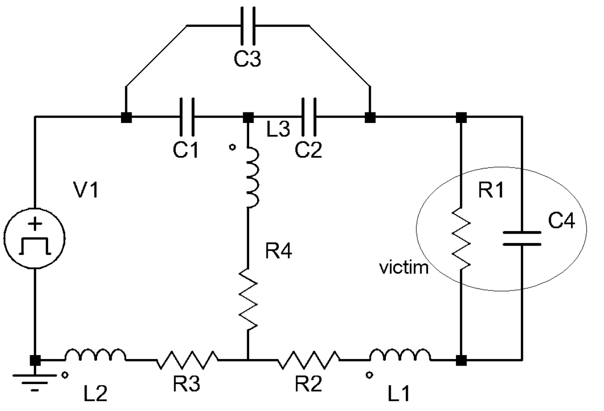
This diagram is a side view, the horizontal line being a solid ground plane, and the vertical line being a solid sheet of metal connected to the ground plane.

There are three reasons for this shield not being completely effective:

- ☹ The field can go around the top of the shield. (EM fields diffract around corners when barrier dimensions are smaller than the wavelength.)
- ☹ The metal of the shield will have a finite conductivity.
- ☹ The connection of the shield to the ground plane will have a finite inductance.

Since this is a simplified diagram of a shield, it is not possible to say which is the dominant weak point, but ordinarily the lack of shield coverage would give the biggest leakage. Next biggest would be the inductance of the connections and finally the conductivity of the screen. As soon as the equivalent circuit is drawn out the situation becomes much clearer.

FIGURE 14.4B:



Notice the definite current paths from the source V1 *and back.*

When modelling with *lumped elements* like this you must always consider the complete current path. The current has to both *go* and *return*. The current does not go into a ground plane and magically 'disappear'.

The shield can be effective surrounding either the source or the victim. Often the source is shielded because there are a lot of possible victim circuit nodes. The shielding will be improved if both source and victim circuits are shielded, but at additional expense.

Shield materials are typically tin-plate or tin-plated brass. They can be ***CNC*** punched or ***chemically milled*** {photo-chemically machined} very inexpensively. For example a sheet of tin plated brass around 30 cm × 40 cm can be chemically milled to as many prototype screens as can be fitted on the sheet for $300 (including the ***NRE***).

The good news is that the metal of the screen is not particularly important when shielding electric fields. Any metal and any thickness will usually do the job, solderability being the deciding factor. The wave impedance is so high that the current flow through the resistive and inductive paths shown as L2, R3, R4 and L3 will not create problems at frequencies below 10 MHz.

The next hardest interference to shield against is magnetic fields below a few tens of kilohertz. High permeability materials such as *radio metal* and *mumetal* are used, but are expensive and heavy. Also, if you make a shield out of mumetal and then drill holes in it, or bend it, you can reduce its relative permeability by as much as a factor of 4×. When a mumetal shield has been formed it is necessary to heat treat it to restore its optimum magnetic characteristics. Traditionally, mumetal screens have been used to shield cathode ray tubes against power frequency magnetic fields and, for sensitive oscilloscope tubes, the earth's magnetic field.

You may think that you should be worried about the magnetic flux density of an interference source and that if it is less than say 1 mT, you need not worry about it. This view is completely wrong. It is not the flux density that causes magnetic induction, it is *the rate of change of the flux density*. Remember your basic physics and Faraday's Law of Induction. In a single loop circuit the induced voltage is given by:

$$\boxed{V = -\frac{d\phi}{dt} = -\frac{d(BA)}{dt} = -A\frac{dB}{dt}}$$

Obviously you need to minimise the loop area, A. However, if B is a sinusoidally varying field, $B = \hat{B} \cdot \sin(\omega t)$, then $V = -\hat{B}A \cdot \omega \cdot \cos(\omega t)$. If the frequency is increased by a factor of 1000, so is the induced voltage. This is why switched-mode power supplies with fundamental frequencies of >30 kHz produce such difficult fields to shield out.

It is possible to shield circuitry from alternating magnetic fields by using a high conductivity shield rather than a high permeability shield. This method is not workable at 100 Hz as the required conductivity is too high. It is a useful method above 1 MHz, but it can still be difficult to implement. In order to do the shielding you need a continuous shield with an impedance below the milliohm level to currents which flow in loops within the shield. The shielding is done by inducing a current flow in the shield, creating an opposing flux.

In the case of the electrostatic shield, a low resistance path was provided for the field to flow down, diverting it from the sensitive circuitry. The field was 'attracted' to the easy path. The same trick works with a high permeability material for a magnetic field. However, electrically conductive magnetic shields are 'repulsive'. The field does not want to go through this path because it is more difficult. The key to success of this method of shielding is therefore to have somewhere else for the flux to go. There needs to be enough space around the shielded area for the flux to be easily diverted.

Shields of this type are very commonly used around transformers at both mains frequency and switched-mode supply frequencies. You may see heavy copper bands wound around the outside of transformers. These are designed to reduce leakage flux from the transformer. The essential point is that they have to be made carefully so that there is a very good electrical bond at the join in the copper band, otherwise they will be ineffective. It is all too easy to add a band that looks ok but doesn't actually work because its resistance is too high.

Shielding against internally generated electric or weak magnetic fields is relatively straightforward compared to shielding complicated electronic assemblies from emitting low power radio signals. You are trying to prevent the equipment from emitting radiation in the range from 150 kHz to 10 GHz, and yet there are bound to be holes in the case for controls, display devices and power feeds. The top-end frequencies that need to be restrained are obviously a function of the clock speeds being used within the equipment.

At frequencies below 30 MHz the RFI emission standards look only for signals coming out of the mains lead {power cord} of a product. These are referred to as *conducted emissions*. The free space wavelength of a 30 MHz signal is 10 m. In order to effectively broadcast such a signal, the size of the antenna would need to be around 5 m. If signal voltages are allowed to escape through the mains lead then these signals will radiate freely using the mains wiring as an antenna. Fortunately these frequencies are relative easy to filter out using a standard mains filter unit.

For frequencies above 30 MHz the RFI emission standards are looking for radiated emissions. As the frequencies increase, the various structures within the equipment can become resonant; seams, joints and holes then become able to let radiation escape. This sort of problem is impossible to solve theoretically with any degree of certainty. Internal wiring only needs to move slightly and the whole internal radiation pattern can be altered.

Some text books give formulae for plane wave shielding factors. Unfortunately they are worse than useless for real-world problems. The term *shielding effectiveness*, SE, is the ratio of emitted field strengths without a shield and with a shield. For a 'thick' metal shield with n holes of diameter d at a wavelength of λ, the plane wave shielding effectiveness is given as:

$$SE = 20 \cdot \log_{10}\left(\frac{\lambda}{2d}\right) - 20 \cdot \log_{10}\sqrt{n} = 20 \cdot \log_{10}\left(\frac{\lambda}{2\mathrm{d}\sqrt{\mathrm{n}}}\right)$$

In this context 'thick' means several skin depths thick so that direct penetration of the shield has a negligible contribution to the loss of shielding. If there is an open area, A, made up n circular holes of diameter d then $A = n\pi \cdot \left(\frac{d}{2}\right)^2 = nd^2 \cdot \frac{\pi}{4}$

The $d\sqrt{n}$ part of the shielding effectiveness formula can therefore be expressed in terms of the area: $d\sqrt{n} = \sqrt{\frac{4A}{\pi}}$

Hence the shielding effectiveness can be re-written as $SE = 20 \cdot \log_{10}\left(\frac{\lambda}{4} \cdot \sqrt{\frac{\pi}{A}}\right)$

… but this result has no term for d in it. Apparently, according to this formula at least, making the holes smaller has no effect on the shielding effectiveness! This result is hopelessly wrong compared to experimental evidence.

There is another correction term to add in to the formula, related to the thickness of the shield, but this term is not enough to correct the errors in the formula. A possible reason for this disagreement between the formula and real life is related to interference patterns. In your elementary physics classes you may recall experiments with water tanks, dippers and strobe lights. These *ripple tank* experiments showed that plane waves passing through small slits produce multiple sources which interact. The resulting interference pattern meant that there were null regions produced.

The same thing happens with electromagnetic waves. A big interference pattern is produced and the tuned dipole antenna averages out the interference pattern, giving a very low reading. Hence as far as measurements are concerned, the emission is very much less with a pattern of small holes. The effect would be that the measured emission level does not decrease as rapidly as it would from a simple point source.

The worst fields to shield against are intense magnetic fields above 100 kHz. Although 7 mm ventilation holes are quite acceptable for shielding against fields up to say 1 GHz, a 500 kHz magnetic field will easily pass through such a hole. Even honeycombed waveguide-beyond-cutoff windows are very poor at shielding these intense magnetic fields. Manufacturers of waveguide windows quote 90 dB or more at >100 MHz, but for near field shielding, using a simple test pickup coil across a honeycomb window, the shielding effectiveness can be as low as 10 dB to 20 dB. Most test equipment will therefore be penetrated by such fields, although the amount of disruption caused will be limited by having small pickup loop areas in the sensitive signal areas. Even expensive gasket materials are very poor at sealing up the equipment: the problem being ***skin depth***.

Skin depth is proportional to the square root of the resistivity of the material. Even expensive silver impregnated silicone gaskets have such high resistivity that the large skin depth requires wide gaskets (say 1 cm) in order to get any reasonable amount of attenuation.

When considering the different effects of magnetic fields and electric fields, it is important to realise that electric fields can produce voltage changes on one wire only. Such a field can be shunted to ground using a large capacitor. A magnetic field, on the other hand, always induces voltage into a complete loop. If there is magnetically induced voltage appearing on a power rail, it will not be improved by decoupling the regulator supplying the power rail. The voltage is associated with the wire rather than the active components. The induced voltage can only be reduced by shielding the loop, or by minimising the loop area.

In microwave amplifiers the shielding itself can cause problems. Consider an amplifier built into an electrically conducting box with a rectangular cross-section. If the width of the box is greater than half a wavelength at any frequency where the amplifier has gain, the possibility of exciting a microwave mode such as TE10 exists. This mode might allow the output to couple back to the input, potentially causing an oscillation. Thus a microwave amplifier which has gain up to 30 GHz should ideally be packaged in a box narrower than 5 mm. Another possibility is to include RAM (Radar Absorbent Material) within the cavity to reduce the feedback. Amplifiers below 1 GHz need only be limited to a width of 150 mm and therefore do not suffer from this waveguide excitation problem.

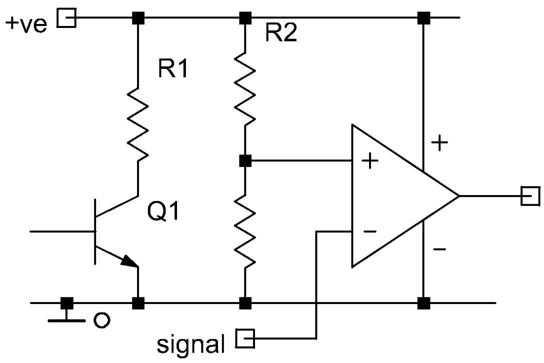
14.5 The Ground Loop

By ground here is meant the signal earth, 0 V, or common line in a measurement or control system. It could also be the mains earth {circuit protective conductor} if that is being used as a common for signals in your system. Ground loops probably cause more problems to analog designers then any other aspect of their work. The reason is that there is nothing to see, nothing to measure, nothing to decouple, and it all seems like a mystic ‘black art’ to the novice.

The world is full of “good advice” as to how to solve the problem of the ground loop. “Use a single point ground”, a *star point*, is the most popular advice. “Experts” love handing out these simple one-line solutions to all your problems, but life is not that simple. To decide how you should implement a “single-point ground at the ADC” in an 8-channel data acquisition system is something that does not get mentioned by these self-titled experts. Only a good understanding of the basics will assist to you to solve these problems, but don’t expect it to be easy!

A circuit diagram is a good way to look at a circuit. The trouble is that not all of the circuit is shown on the circuit diagram. The circuit diagram does not show ***strays*** for example. When trying to eliminate noise, interference and ***cross-talk*** problems it is important to realise that strays and ***parasitics*** exist. It is also important to be able to draw them onto the circuit diagram. Indeed the circuit diagram can be deliberately drawn to emphasise the problematic elements of the design.

FIGURE 14.5A:

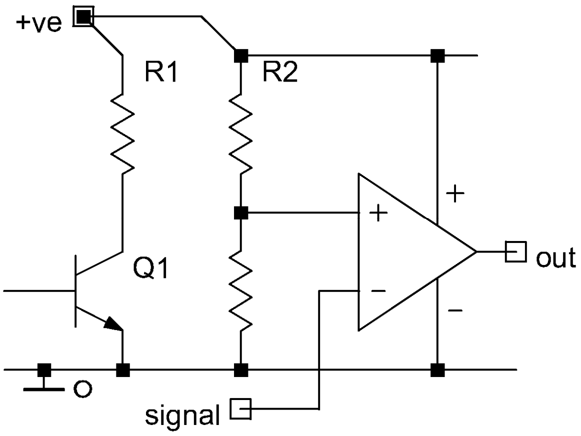


In this partial circuit, Q1 is being switched, causing a current to pulse in R1 and hence in the +ve {positive} rail. It is obvious that the +ve rail has a source impedance, and that this pulsing current causes a pulsing volt drop on the +ve rail. The circuit board track resistance might not be large, but on sensitive amplifiers this sort of problem needs to be avoided.

EX 14.5.1: In the above circuit, assume that the tracking was done to follow how the circuit was drawn. Neglecting the source resistance of the positive rail, what is the change in potential at the top of R2 when the transistor is switching a 10 mA load. Take the track resistance as 1 mΩ/cm of length and each track segment as 5 cm long.

FIGURE 14.5B:

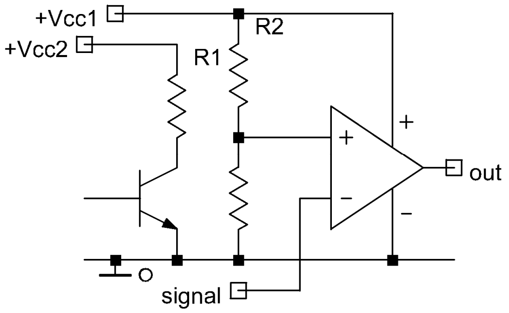
This unusual re-drawing of the circuit reminds you and the person doing the PCB layout to route the traces separately to R1 and R2, thereby minimising the common resistance path.



The exercise above was simple because I told you exactly what to look for. If you just had a pulsing voltage on the amplifier output, you wouldn't have known where to look. Was the noise source due to magnetic induction, capacitive coupling, resistive volt drop, or something else? The coupling mechanisms are simple and obvious when you draw them on an equivalent circuit, but when you have only a circuit diagram and a PCB layout, it looks quite a lot more mysterious than that exercise would make you think.

It would be much nicer if a little genie opened up a text book for you at the right page when a noise problem developed. Since little genies don't exist, it is up to you to first of all establish which *type* of problem you have in front of you; you can then 'open up' the relevant section of your experience or text book.

FIGURE 14.5C:



It is certain that you will not get a +ve supply with a source resistance <1 mΩ. In this case you might need a separate supply in order to remove the regulation effect.

Suppose the current is pulsating sufficiently slowly that you can see the voltage changes with a DMM. You can still measure the voltage on Vcc1 changing; but why, it is running from a separate regulator?

In desperation you may even wire up a bench supply or a battery to get a 'clean' {noise free} power rail. This is where a simple fact can save you hours, days or even weeks of trouble trying to find the elusive coupling mechanism. It is quite possible that you have reduced the pulsations on R2 by a factor of 100× doing all this isolation of power supplies, but you are still annoyed that you can see the effect at all.

What is less obvious is that the ground path also has an impedance. Even if it is a solid ground plane, it will still have an impedance. Don't think that just because you have a ground plane you do not need to worry about ground loops. The effect is reduced, not eliminated. It is then a matter of the sensitivity of your circuit.

In this case you would find that there is still a common resistance path for the pulsing current and the current going to the resistive divider chain. Note the remarkably confusing word 'common' used there. In this context it means "shared by several". The common resistance path just happens to be in the ground path [also known as the *common*!].

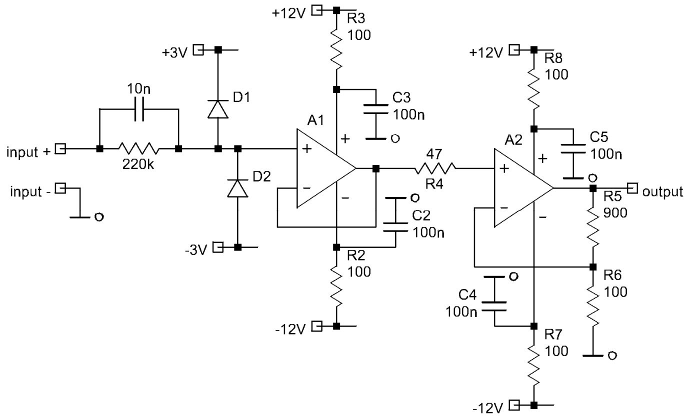
So there you have it! A ground loop is a problem due to the fact that currents flow through pieces of wire or PCB tracks and generate small but significant voltages. And yet this simple layout problem causes more trouble than you can possibly imagine.

In precision LF measurement systems (<0.01%) it was usual to designate one spot in the instrument as a 'star point'. Multiple signal wires were taken from this point to give each sub-circuit of the overall system its own 0 V reference point. This technique was and is quite workable, but is perhaps not as fashionable as it once was. Nevertheless local star points are an important means of making accurate measurement circuits.

A single star point is not workable in measuring instruments for frequencies >10 MHz. In this case there may be too many high speed interconnects and it is better to use the chassis as the reference. The key thing to remember about the routing of the high speed lines is that current has to both *go* and *return*. It is very poor practice to route digital bus lines down ribbon cables and then let the RF signal currents find their own way back to the driving PCB. It is best to route ground wires {0 V} next to the signals, minimising the magnetic loop area. This makes the antenna loop-area less, both as a transmitter and as a receiver. This 'loop-area' idea is so important that it is possible to buy ribbon cable with the wires twisted in pairs to minimise it.

Rather than use a ground connection between every signal, it is possible to use a ground for every two signal wires. This way there is always a ground wire on one side of each signal wire. This provides the necessary RF return path and keeps the loop area relatively low. Reducing the number of ground wires in this way is not ideal though, as it still allows crosstalk between the adjacent signal lines. However, it may be a useful compromise between the width of the cable and the resulting emission/pickup problems.

FIGURE 14.5D:



This is a design for a preamplifier with a high impedance, protected input, and a gain of ×10. With the inputs shorted it is found to be too noisy. Inspection of the output shows a non-random aspect to the noise. This is evidently due to system interference, rather than the random noise of the components.

To 'kill off' the noise, do not try just one thing at a time; just increasing C5 may not show any improvement, as the effect could be 'swamped' by other sources. Reduce the circuit down to a minimal configuration which is noise-free, then you can see each little bit of noise contributed by each section. The first step is to remove R4 and short the input of A2 to the bottom of R6. Now I don't just mean connect it to 0 V; I mean get a piece of wire and solder one end to A2 non-inverting input and the other end to the lead/pin of R6.

If this configuration is not quiet enough then try increasing C4 and C5. Have some big tantalum capacitors (22 µF) and low-ESR electrolytics to hand when doing this sort of work. I have made it easy for you by including R4. You probably won't be so lucky and you may have to cut tracks or lift component legs to do this sort of thing. Too bad; if you worry about cutting tracks on the board then you will never get the job done! Noise problems may require extra holes in the chassis, extra screens, more screws; all very mechanical. Be proficient with a hacksaw, a file, a drill, emery cloth &c, and don't be afraid to get your hands dirty. You will not be able to fix all of your problems using a SPICE simulation!

Now the sub-circuit is quiet you can go back to include the previous stage. Short the input of A1 back to the bottom of R6 and see if the amplifier is still quiet. If not then try C2 and C3. Now link the +input to R6. Remember to use as short a piece of wire as you can. The wire length may be getting long now and this path is therefore susceptible to magnetic and electric fields. You may need to put a shielding lid in place to stop stray pickup giving you a falsely noisy signal.

If it is noisier now then you may suspect a path through D1 and D2. A noisy supply to the clamp diodes can couple through the diode capacitance onto a high impedance input. Let's suppose it has been quiet all the way through up to this point. Then as soon as you connect a short directly across the main input terminals the noise comes back. The answer is that the input ground is the –input terminal and the reference ground for the amplifier is the point at the bottom of R6. This is why I have been emphasising that your wire link must go to the bottom of R6. Now you have discovered the ground loop for yourself.

To fix this problem you must disconnect R6 from its existing ground connection and run a separate wire from R6 back to the –input terminal. You will improve things by adding screens which reduce the impedance from R6 back to the –input terminal, but a parallel path will not kill off the noise entirely, it will simply attenuate it. If this level of attenuation is sufficient then leave it there.

If you need more attenuation of the interfering source, then you have to start splitting the ground paths up so that less noise current is allowed to flow in the wire link from the end of R6 back to the –input terminal. This is a difficult problem because you have to imagine where the currents are flowing and guide the noisy ones away from the signal paths. This can be done by using separate ground planes (expensive), by selective cuts in the ground planes, or by the use of local star points. I can't give you a general fix for this; you need to understand the basics. Once you see that the ground currents are causing the problem you will be able to work out how to make them flow correctly. This is only Ohm's law after all!

On a single-channel, low frequency, high resolution system, a star point grounding scheme is very useful. A 1 mA current change flowing in a 1 mΩ common resistance creates a 1 µV signal, which may be entirely unacceptable. Thus a ground plane system may not work.

On a multi-channel, high frequency, lower resolution system, there are too many signals flowing here, there, and everywhere, to constrain them to a star point system. Rather than travel back to the star point, the higher frequency signals will tend to flow capacitively back by the shortest paths. Ground planes are now essential. However the inductance of the ground plane will mean that there are appreciable voltage differences between different parts of the ground plane. Signals should now be routed differentially if possible. Even if the circuit appears to have a single-ended input, you must consider the ground as one of the two signal wires. Thus you may decide to have a ground plane and also a separate signal ground wire used for the other half of your differential input signal.

Consider a control line coming into a sensitive preamp. If you decouple this line to the ground plane in the preamp, any voltage noise on the control line will have been short-circuited to ground. You have fixed the noise problem, since capacitive coupling from the control line to the preamp has been eliminated. True, BUT you have now injected all the picked-up signal as a current into the ground plane. Any fast edges or high frequency content will therefore create voltage drops across the plane. You can actually make the noise worse by this inadvisable technique.

The answer is to always use an RC filter rather than just a capacitor. The resistor dramatically reduces the amount of *current* injected into the ground plane and also improves the filtering. Control lines are often long and will inevitably pick up all sorts of noisy signals on their way through the system. You would do well to assume that they have such noise on them until proven otherwise. Thus in a system of any complexity, it is hardly worth proving that these RC filters are necessary; just add them anyway to reduce design and debug time.

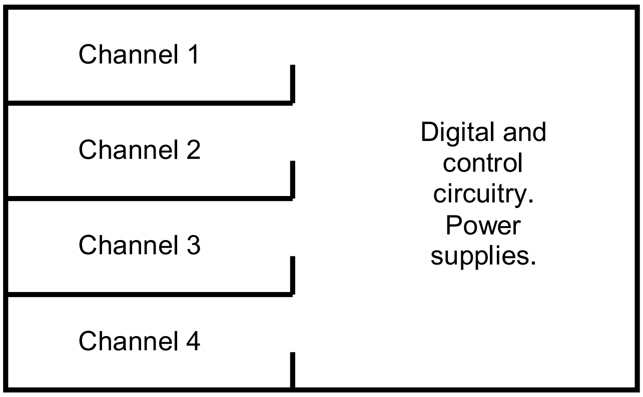
Sometimes you need to use star point grounding for DC accuracy, but there are fast transients being passed down these highly inductive paths. In this case it may be advisable to also have a ground plane. The ground plane can be used for RF decoupling of the DC ground paths which all individually route back to a single star point. This grounding scheme then behaves like a single point ground at DC and a ground plane system at higher frequencies.[1]

Much more emphasis is placed on the structure of the ground plane compared to the

[1] H.W. Ott, 'Chapter 3, Grounding', in *Noise Reduction Techniques in Electronic Systems*, 2nd edn (Wiley, 1988).

power plane(s). The reason is that a noisy power rail can be filtered with respect to a quiet ground plane; a noisy ground plane cannot be filtered.

FIGURE 14.5E:



Ground plane layout for low-noise multi-channel preamp. The lines within the drawing represent cuts in the ground plane *and* routing keep-outs on other layers. This layout minimises inter-channel cross-talk. This is a real layout from four channel 20 MHz 12-bit scope preamps, as well as 8-bit 500 MHz scope designs.

The rules for the ground plane design process are:

- Cut between channels in order to minimise cross-talk.
- Route control and power signals in the area where there is no cut in the ground plane.
- No tracks or power planes should cross a cut in the ground plane.
- Use a solid plane without any regard for "copper balancing".
- Do not use cross-hatched power planes. (If the PCB vendor objects, get a better vendor.)
- Use cuts to allow currents in switched-mode supplies to be contained within their own tight little areas.
- Use cuts around oscillator circuits so that phase noise is not increased.
- Remember that current flows in closed loops. Make sure that any noisy track has a loop path which does not pass through sensitive circuit areas.
- Further segmentation of the ground plane in the channel areas may be necessary for optimum performance.
- If there are ADCs then cuts in the plane in the area of the ADC will be necessary to reduce the amount of current that flows from the digital circuitry back into the analog circuitry.
- Cuts in the ground plane should not be <0.020″. There will otherwise be capacitive coupling across the gap. A gap of between 1 mm (0.040″) and 3 mm (0.120″) is preferable.
- Do not allow heavy current circuitry anywhere near this ground plane system. Motors, contactors, heavy duty relays and other power switching components should have their own separate grounding scheme to minimise common current paths.
- Do no put sensitive components or tracks near the end of a ground plane cut. The cut forces any HF current to get confined to a narrow region and the result is a noticeably increased magnetic field just in that area.
- Be willing to bend or break any of the above rules if they make the design too difficult. It may still work well enough.

- When you put cuts in a ground plane, have large exposed pads over the cut, connected to the ground plane by multiple vias, so you can try remaking the cut to see if the noise is improved. This gives you a fall-back position, allowing fine tuning without the necessity of re-laying the PCB to see if it works. It also allows experimentation to show if the cuts proposed are useful or not.

You may be working on such a cost sensitive project that you cannot afford a whole routing layer devoted to a ground plane. In this case try to link the ground connections to components by multiple separated traces in a sort of grid network. This gridded-ground will considerably reduce the inductance of the ground connections and improve the result.

14.6 The Switched-Mode Power Supply

Efficiency is often a major concern when designing a switched-mode power supply (SMPS), so there has been a lot of study into improving the efficiency of the magnetic components. It is not my intention here to list and comment on all types of switched-mode power supplies. I merely wish to give you some hints as to their use. Firstly you should know why you would want to use one at all.

Electronics goes through fashions {fads; public relation trends} to a lesser degree than other fields, but there is still an unacceptable element of 'fashionable' design. Around 2003 it became 'fashionable' to have micro-processor controlled domestic appliances. Regardless of whether or not they performed better, they were sold to the public as the latest greatest invention.

This is an easy trap to fall into, and with domestic equipment being sold to the consumer market, you may not have any choice but to follow it. If you have the choice, then please use intelligence rather than fashion to guide your designs.

Switched-mode power supplies are great. They do lots of nice things. They also have drawbacks and they may not be suitable for your application. There are two contenders: a switched-mode power supply and a linear supply. Let's compare them on a point by point basis.

Efficiency

Don't believe the hype {marketing propaganda}. Whilst a linear supply can be less efficient than a switched-mode supply, it can also be more efficient! It really depends on the situation. If you want a piece of equipment to run from a mains supply of 180 V to 265 V then a switched-mode supply can do that easily, and its efficiency can be fairly constant over that range.

For international use, 180 V to 265 V is a standard range of values for the "230 V" range of a piece of equipment. Some types may even cover the whole international range from 90 V to 265 V in one go. Others would have a single switched range. A transformer and linear regulator design would be *horribly inefficient* in this application as will now be demonstrated.

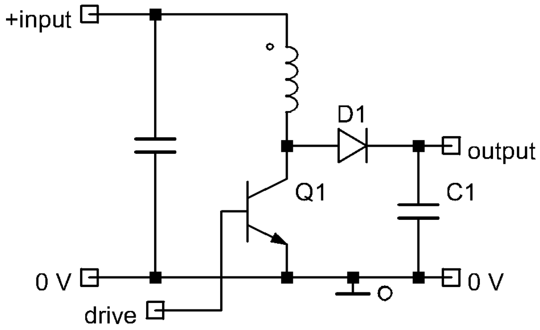
EX 14.6.1: A transformer and linear regulator is used to generate a +15 V rail at 1 A. This works from 180 V to 265 V. Estimate the best achievable efficiency at 265 V if there is no tap-changing done on the transformer.

One alternative is to use a *ferro-resonant transformer*, also known as a *constant voltage* *transformer*. This combines the low-noise of a linear supply with the wide input range of the switched-mode supply.

Noise

Strictly speaking the noise I am talking about should be called interference, but in common usage engineers and users just call it *noise*. A linear supply is quiet. There is no switching going on; it just sits there quietly getting hot and doing little else. A switched-mode supply, on the other hand, is switching vigorously between say 10 kHz and 1 MHz. It is generating switching spikes as conducted power & ground spikes, 'radiated' electric fields and 'radiated' magnetic fields. In fact switched-mode supplies can generate so much noise that they often need to be put in their own screened boxes, well away from the sensitive circuitry. If you understand the origins of the noise then you stand a chance of reducing it.

FIGURE 14.6A:



This is a *boost* converter. The output voltage is always larger than the input voltage. The switching device has been shown as an NPN bipolar transistor, but it would usually be a MOSFET or part of the controller chip. It could even be a mechanical switch.

The operation is very straightforward. Q1 is turned fully on and current starts ramping up in the inductor. When the transistor is turned off, the current in the inductor keeps on flowing in the same direction, with the same magnitude, and therefore flows through D1 into C1. This current decreases as the magnetic field [the energy stored] in the inductor decreases.

In all switched-mode power supplies there is one essential equation that you need to have at your fingertips. You must know this by heart [memorise it] and you will be able to handle all types of switched-mode power supplies.

$$\boxed{E = L \cdot \frac{di}{dt}}$$

If you dislike simple calculus expressions like this one then you had better get out your elementary texts and have another go at them. You absolutely need to be comfortable with that expression. When I said the current in the inductor ramped up, I meant that literally. From the equation it should be clear that a constant applied voltage gives rise to a constant rate-of-change of current. The current waveform is therefore a linear ramp.

Looking at the boost converter circuit, you should be wondering what sets the output voltage. The answer is the feedback loop [which has not been shown]. The ***duty cycle*** of the drive waveform is adjusted by the control loop to give the required output voltage.

This next equation is the key to solving switched-mode power supply problems:

$$\boxed{\text{Efficiency, } \eta = \frac{\text{Power Out}}{\text{Power In}} = \frac{\text{Power Out}}{\text{Power Out} + \text{Losses}} = \frac{\text{Power In - Losses}}{\text{Power In}}}$$

EX 14.6.2: A switched-mode supply is 75% efficient and is driving a load of 234 W.

a) What is the power dissipated in the supply?
b) What is the input power?

The efficiency equation can be usefully re-arranged to evaluate how much power has to be dissipated with a given efficiency and power requirement. This affects the cooling inside an instrument.

$$\boxed{\text{Losses} = (1-\eta)\times \text{Power In} = \left(\frac{1}{\eta}-1\right)\times \text{Power Out}}$$

To understand the boost converter better, look at the problem of the circuit malfunctioning.

EX 14.6.3:

a) What happens if Q1 fails short-circuit?
b) What happens if Q1 fails open-circuit?
c) What happens if the control loop breaks and Q1 is switched on permanently?
d) What happens if the control-loop feedback resistor goes open-circuit?

I was once asked to give an opinion on a very similar supply to this one, used as a small part of an existing design. The transistor kept failing and it was noticed that the inductor smoked visibly at power on! It turned out that the control circuitry took several milliseconds to start up, and during this time it unintentionally had the transistor turned on. This is a very common type of problem and is easily solved by AC coupling the base/gate drive to the switching device. In any case, it is very bad practice to allow a single-fault to cause multiple component failures. AC coupling the base drive in this circuit only costs one small capacitor, perhaps $0.04. This is well worthwhile.

Now that you have looked at the circuit a bit, let me state that there are two distinct modes of normal operation of this circuit which relate to the control loop and the load rather than to the circuit as drawn. The modes are *continuous* and *discontinuous*, relating to the current in the inductor.

The discontinuous mode is the easiest to understand. It means that the inductor current stops {it drops to zero} before the end of the cycle; it is not continuous. If it were a capacitor, you would say that it was fully discharged (or uncharged) at the end of every cycle, but there is no equivalent term for an un-energised inductor.

The cycle starts by turning on the transistor. It is turned on rapidly and totally. The current in the inductor increases linearly, if it is a well behaved inductor. At some point the transistor is turned off. If this turn-off point is set by sensing the inductor current this is *current-mode control*. If the switch-off is based on the duty cycle, which is in turn set by the feedback, this is *pulse width modulation control*. Note that both control systems change the duty cycle.

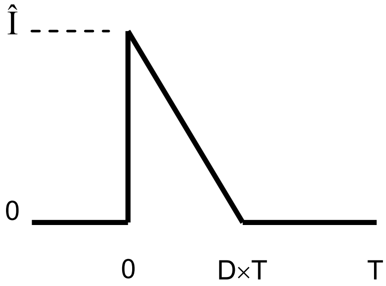
When the transistor has been turned off, the current in the inductor has to go somewhere. It goes through the diode into the capacitor C2. C2 charges up until the inductor runs out of energy.

***EX 14.6.4:** Neglect all losses unless they are explicitly mentioned. Neglect all tolerances. This exercise is to give an overview of the function of a switched-mode supply. The input voltage is 10 V. The output voltage is 100 V. Take the diode drop as 0.7 V. Assume the inductor is pure (no resistance) and linear (it does not saturate with the currents and loads given). Calculate on the basis of *discontinuous* mode operation.

a) What is the ratio of the times of the inductor being 'charged' to the inductor being 'discharged'? (One could say the magnetic field is being 'charged'.)
b) The loop is running at a fixed frequency of 25 kHz. The maximum output load that the circuit can supply is 75 W. What is the value of the inductor?
c) Draw the ideal current waveforms in the inductor and the diode.

The current waveform in the diode is typical of waveforms found in SMPS. It is therefore important to know the relationship amongst the peak, mean and RMS values of the waveform, according to the duty cycle of the triangle.

FIGURE 14.6B:



EX 14.6.5: This is one cycle of a repetitive current waveform, period T. Derive expressions for the mean and RMS currents.

@EX 14.6.6: What are the RMS values of the following waveforms?

a) A rectangular voltage waveform with a low level of zero and a high level of V, having a duty cycle of D.
b) A rectangular voltage waveform with a low level of -V and a high level of +V, having a duty cycle of D.
c) A rectangular voltage waveform with a low level of -sV and a high level of +V, having a duty cycle of D.
d) A rectangular voltage waveform with a low level of +sV and a high level of +V, having a duty cycle of D.
e) A rectangular voltage waveform with a low level of zero and a high level of V, having a duty cycle of D, when AC coupled.

EX 14.6.7: What is the worst uncertainty on the measured RMS value of a rectangular waveform (low value= zero, high value= V, duty cycle= D) using ADC data over a non-integer number of cycles (>1). Neglect the ADC resolution, ADC conversion accuracy and time resolution.

Trying to directly measure the losses in SMPS components is remarkably difficult.

Consider the diode in the boost converter, shown previously. You could measure the power loss in this diode by monitoring the current and voltage waveforms simultaneously on a scope, or some sort of power meter. The problem is that the forward drop of the diode will be from 0.4 V to 1 V, depending on the type of diode, but with 100 V output and 10 V input the reverse voltage will be up to 90 V. When the voltage is –90 V, the current will be small, but the measuring device still has to function. The dynamic range needed from the measuring device is therefore very high. If you want to measure the forward drop to 1% accuracy, that is 1% of 0.4 V which is 4 mV. But the measuring device needs to measure 90 V as well, so that is a resolution of 4 mV in 90 V, or 1 part in 22,500. In ADC terms this would require 15 bits or more.

Now that was just to measure the voltage; you still have to measure the current. There is a choice of a current probe or a small series resistor. Obviously the small series resistor has to be made small enough to have minimal effect on the reading. All this is very unpleasant and the fact is that power losses in diodes, transistors, inductors and capacitors are seldom measured directly.

You can just use the formula

Power Out = Power In – Losses

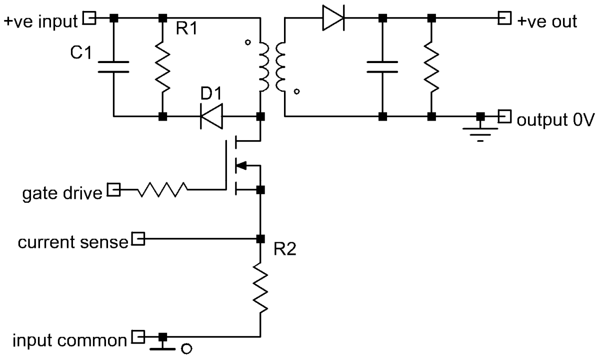
to calculate the losses. If a new component is used you can see the improvement directly by the change of "power-in minus power-out". When it is desirable to find out which components are responsible for the majority of power loss, you can use a *thermal transfer* technique. If the device is dissipating power it will get hot. The switched power it is dissipating is very hard to measure, but if you can take the *same* device, on the same heatsink, and heat it to the *same* temperature using DC power, which you can easily measure, you then know what the switched power is. Obviously this works best when the temperature rise is large. In fact if the temperature rise is not large then you can make it large by deliberately (thermally) insulating the component being tested. This makes the temperature rise greater and therefore makes the transfer measurement more accurate.

A point mentioned in a previous exercise was that for buck and boost converters, power flows directly from input to output during part of the cycle. This should be contrasted with a flyback scheme using a transformer, where all output power comes from stored energy in the inductor. The lower stored energy in the inductor means a smaller size and the possibility of greater efficiency.

Current-mode control

I mentioned current-mode control previously without much emphasis. It is an important subject. One of the main difficulties with switched-mode supplies is *primary ripple rejection*.

FIGURE 14.6C:



This is a flyback converter. D1, C1 and R1 form a ***snubber*** network to stop the MOSFET from getting blown-up by over-voltage. For simplicity I have not shown the controller chip or the feedback path from the output back to the gate drive.

Ignore the current sense resistor (R2), for now; just pretend it is vanishingly small. Suppose the positive input supply has ripple on it because it is being driven from a rectified mains supply, with or without a transformer. This power-frequency ripple is at say 100 Hz [or 120 Hz if the supply frequency is 60 Hz]. The switched-mode supply may be running at 50 kHz. If the duty cycle is held constant, this ripple will pass straight through the switched-mode supply and will appear at the output. The feedback loop tries to eliminate this ripple.

The feedback loop has a finite gain and therefore a finite primary ripple rejection. You can't just increase the gain because the loop will go unstable. Furthermore, it is nice to be able to have lots of ripple at the input side of the switcher.

More input ripple means:

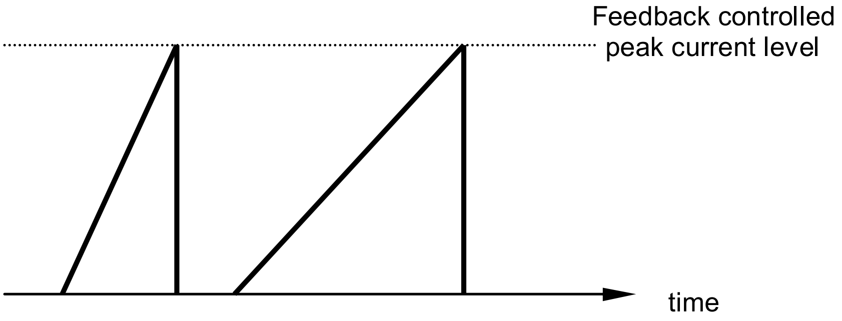
☺ Smaller input capacitors [meaning lower cost]
☺ Higher *conduction angle* in the input diodes [Higher efficiency]
☺ Better power factor [lower harmonic currents drawn]

If you have 20 V ripple at the input and want 1 mV of ripple [at 100 Hz] at the output then this is a ripple rejection of 20,000 or 86 dB. This is not going to be achievable with a direct pulse-width modulation control.

A very elegant solution is a current-mode controller. These are available as 8-pin chips for $1 so you would just use one of them to solve the problem. The idea is very simple. The inductor current ramps up and when it reaches a predefined level the controller shuts off the MOSFET. It is not the duty cycle that is being adjusted [directly], it is this peak current in the inductor.

The feedback circuit sets a desired peak current in the inductor. With a constant load and a constant switcher frequency, this peak current will be constant. Since the MOSFET is switched off at this peak current value, the primary ripple does not have any effect on the feedback loop. Primary ripple rejection is remarkably high without the main feedback loop being involved. Any remaining ripple will be further reduced by the main feedback loop. Such a scheme would have no trouble achieving the ripple rejection suggested in the previous paragraph.

FIGURE 14.6D:



The effect of a reduced input voltage is a slowed rate of rise of inductor current, but to the *same* peak level.

You are paying for improved primary ripple rejection by having a slightly more complicated control chip. There is also power loss in the current-sense resistor. Once you have figured out what the controller is doing, you can eliminate the current-mode sense! After all, the current sense is simply a ramp related to the input voltage. The input voltage could be sensed directly and used to modify the duty cycle more efficiently.

Power Factor Control

Now the current-mode approach is designed to maintain a constant power output and consequently draws a constant power from the power source. This is not ideal for the power company however. They prefer it if all loads look resistive, since harmonic currents dissipate power which is not being paid for by the customer.

If a switched-mode supply is going to behave like a resistive load then it needs to draw power in a very inconvenient manner as far as the electronics are concerned.

***EX 14.6.8:** Describe mathematically the power drawn by a resistive load as a function of time for a sinusoidal input voltage.

You can buy controller chips which provide this very function and are referred to as *power factor controller* chips. You should understand that if power is being drawn from the mains supply in this sinusoidal manner, but it is being consumed at a constant rate, then there will have to be some compromise somewhere; you get ripple.

There are several ways to overcome this problem:

- ☺ Don’t specify the power frequency ripple at too small a value.
- ☺ Use huge output capacitors.
- ☺ Use a two-stage switcher.

This last solution is the usual approach. A power factor corrected ‘front end’ is added to an existing design to improve the power factor without excessive power frequency ripple on the output.

One of the key applications for power switching devices is off-line {mains} switched-mode power supplies. The best advice I can give you on these is to avoid designing them as part of an overall system design. Buy in an off–the-shelf design or get a design customised for your application. The designs look simple and they are even drawn out in manufacturer’s application notes. Don’t be fooled! This is an easy trap to fall into, and the design time can end up being months and months.

It is easy to get the first prototypes working, but to get the design to be robust against mains transients and start-up conditions will take considerable time. It is also very demoralising because just when you think the design is working, one unit ‘blows up’ for no apparent reason several months later. You can never say that the design is "fully working" as you would have to test it indefinitely; all that can happen is a unit can fail. Unless you are going to be shipping these off-line switchers by the thousands per year, the time and the resources required to get the design working well will not be well spent.

You have to think about start-up and shut down conditions. The usual 'trick' with switched-mode supplies of any description is that a fault can cause them to blow up the active switching device(s). You replace the switching device and it just blows again. You therefore have to work out some scheme to get the system running in order to see what is wrong with it. The classic fault is that the clamp circuit fails. This puts over-voltage on the switching device which then fails. Until you fix the clamp circuit, you can't replace the switch without having it fail again. Of course you can't tell that the clamp has failed until you run it and see the transients!

This problem is so bad that commercial repairers of TVs and videos just replace all the parts in the switched-mode supply if there is a fault. Vendors even sell kits of these parts, knowing that it is cheaper to replace the whole lot rather than trying to fault-find the circuit!

Designers can't follow this technique though. We need to find out why the circuit is failing and rectify {correct} the problem. The SEEKret in this case is to run the system from a reduced voltage supply using external bench power supplies. This is particularly true with off-line switchers. There is often a start-up system consisting of a capacitor and an auxiliary winding. When the system receives power the capacitor charges slowly via a large valued resistor. As the voltage reaches a trip point the switcher starts, and as soon as it starts to switch, it generates a power rail for itself off of an auxiliary winding. Now if there is any fault anywhere in the circuit, it won't generate its own supply and therefore it won't run. The trouble is you can't see what is wrong because there is no power on the circuit now! The bench power supply circumvents the start-up problem and will help to diagnose the fault. This sort of circuit is very simple and cheap, but the fault finding is inevitably far more difficult and dangerous than with ordinary circuitry.

Protective Measures

It is not very sensible to include a switched-mode supply directly on a circuit board, powering expensive circuitry. When the board is first made, if a component has been misplaced, or if a solder joint is faulty, you could end up spending a lot of time and money fixing the resulting destruction.

Suppose you have a buck converter producing 2 V from a 12 V rail. If the switcher is not working properly it might produce 12 V. Anything connected to this power rail would get destroyed. Obviously you have to weigh up the cost of the protection circuitry against the cost of replacing the destroyed circuitry, taking into account some estimate of how often the switcher would fail or be built incorrectly. This last factor of the probability that it could go wrong, or be built wrong, is only really going to be found out by experience.

The typical solution to this problem would be to design in a ***crowbar*** activated by some defined over-voltage event. If you don't wish to go to the extra expense of fitting a crowbar then I would strongly recommend that as part of the normal manufacturing process, the switcher part of the circuit should be powered up on a dummy load before being connected to the main circuitry. This should minimise re-work costs.

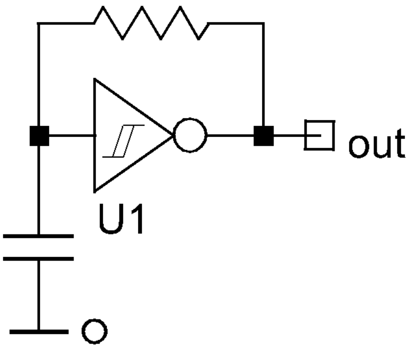
It is not just switched-mode supplies that suffer from this faulty build problem. You can get exactly the same problem with linear regulators. Although off the shelf regulators are extremely reliable and unlikely to fail, the "universal" type use two resistors to set the output voltage. If one of these resistors is open circuit by reason of a dry joint, or if the wrong resistor value is fitted, the load circuit could easily get destroyed by the resulting over-voltage. Another useful trick in this case would be to use a large zener diode from the power rail to ground. If there were a problem, the zener could shunt the excess current to ground, preventing a major catastrophe. Just make sure the zener has enough power handling capability for the resulting continuous overload current.

14.7 The Oscillator

Any idiot can make an oscillator; wire up an amplifier badly and you can get one. Let me re-phrase the original statement; any idiot can get an *accidental* oscillation. If an amplifier is wired up so badly that it oscillates, that is not an oscillator design. When you design an oscillator, you are trying to make both the frequency and the amplitude of oscillation stable. You also need to make these parameters repeatable from unit to unit. Having a spurious feedback path around an amplifier gives an unstable, unpredictable, and unrepeatable oscillation.

The term 'oscillator' covers a vast range of devices and techniques, but there are fundamental similarities between some of the types. One of the simplest electrical oscillators is the ***Schmitt Trigger*** oscillator.

FIGURE 14.7A:

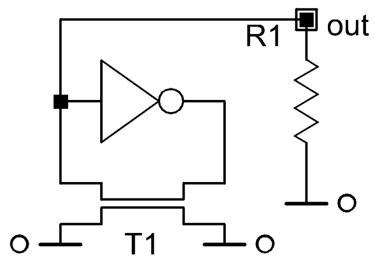


For this circuit to work correctly, the gate U1 must have a ***Schmitt*** *input*, meaning that it has hysteresis on its switching levels. If the input is rising the output may go low for an input of 3 V, whereas when the input is falling the output might not go high until a level of 2 V is reached.

Packages of CMOS Schmitt inverters {six per package} are cheaply available [$0.10], but the tolerance of the thresholds is very poor, meaning that the frequency of oscillation is not well defined. Nevertheless they do make very inexpensive and reliable oscillators.

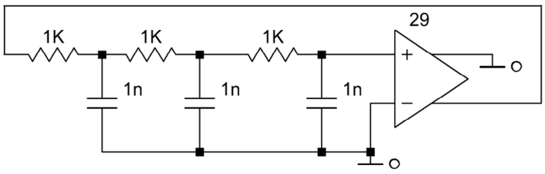
EX 14.7.1: A CMOS Schmitt inverter, as above, runs from a positive supply V_P and has switching thresholds of V_L and V_H. Neglect the propagation delay in the gate. Derive an expression for the frequency of oscillation, stating any simplifying assumptions made.

FIGURE 14.7B:



R1 is the load of the next stage and T1 is a transmission line, which could be just a long piece of wire. This oscillator design relies on the propagation delay through the gate and through the transmission line to function. This design is only useful for >10 MHz applications since the transmission line will be large and expensive for delays longer than a few tens of nano-seconds.

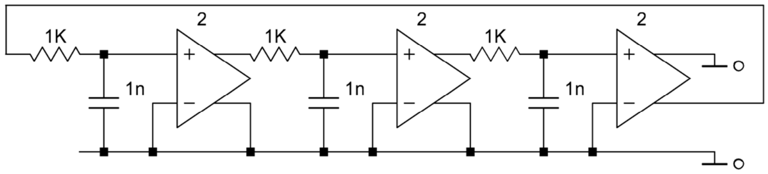
FIGURE 14.7C:



The phase-shift oscillator, consisting of an inverting amplifier and 3 identical unbuffered RC stages dates back to 1923,[2] but is not a good starting point for an introduction. The complexity of the network calculation[3] hides the simplicity of the circuit.

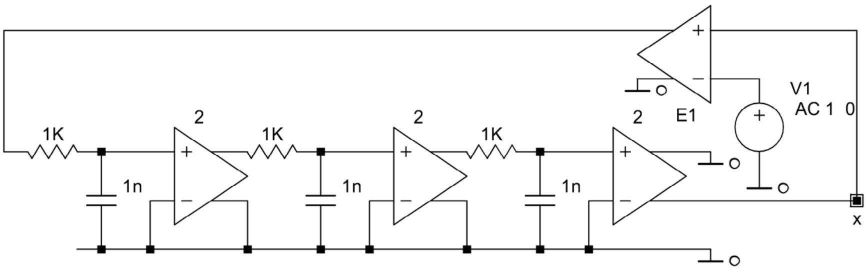
In elementary texts you are given a phase-shift network or a resonant network and you work out the oscillation frequency. That is all very well, but the real problem is one of amplitude. Here is a simplified "text book" phase-shift oscillator:

FIGURE 14.7D:



The phase shift of each RC network is designed to be 60°, and since the attenuation at this phase shift is 2, the buffer amplifiers also have a gain of 2. The loop-gain is 1 at the oscillation point and you have a stable oscillator. The only problem is that if you make this as drawn, it either doesn't oscillate or it produces a distorted output signal. You can simulate the above circuit on a SPICE transient analysis and you will get *nothing* out. To investigate why, a second circuit can be tried. This has the same gain around the loop, but it allows the injection of an external signal.

FIGURE 14.7E:



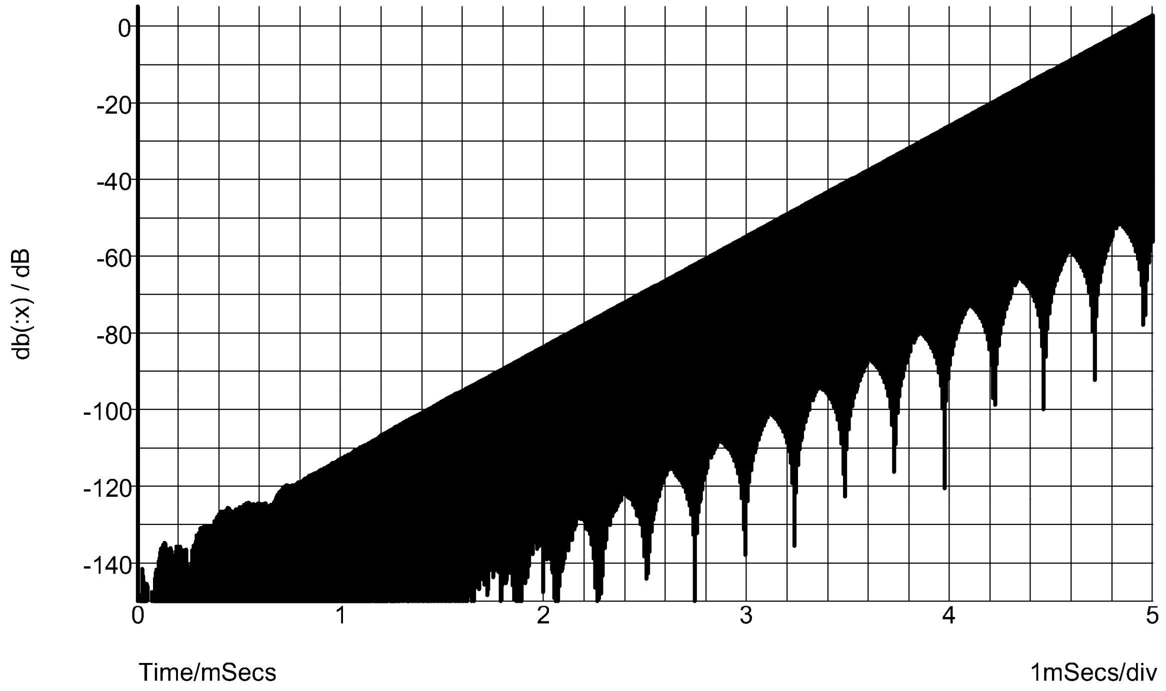
A SPICE small-signal AC analysis shows a resonant peak at the expected frequency of 275.66 kHz, but the amplitude and Q look as if they are heading towards infinity.

If source V1 is replaced by a noise generator and the gain of amplifier E1 is increased to 1.01 then a transient analysis does show a response.

[2] H.W. Nichols, 'Reamplifying System', *US Patent 1,442,781* (Jan 1923).

[3] E.L. Ginzton, and L.M. Hollingsworth, 'Phase-Shift Oscillators', in *Proceedings of the IRE*, 29 (Feb 1941), pp. 43-49.

FIGURE 14.7F:



The oscillation frequency signal is amplified on each pass through the loop by 1.01 (1% or 0.086 dB), giving a geometric increase (linear on a log scale). Each pass through the loop is a minimum of one oscillation frequency cycle. Therefore the start-up time in cycles is governed by the initial noise, the final limit value, and the loop-gain.

$$\boxed{\text{Start Up Time} \geq \frac{1}{f_{OSC}} \cdot \frac{\text{Final Output (dBV) - Initial Input (dBV)}}{\text{Loop Gain (dB)}}}$$

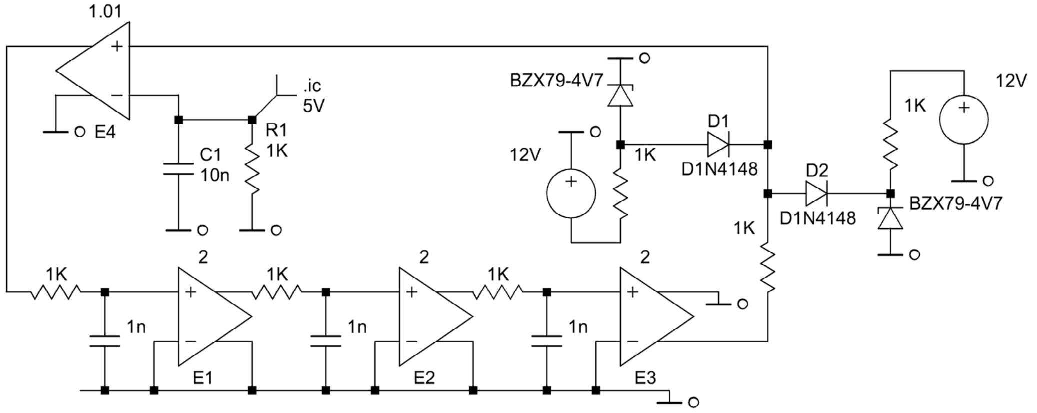
There is no special reason for using dBV. It is ok to use any dB measurement of the input-to-output values as it is only the dB ratio that is important. It is clear that an oscillator can be made to start-up more quickly by either increasing the *excess loop-gain* or by 'kicking it' hard, rather than allowing the noise to make it start.

Obviously in this simulation the output increases indefinitely. When the amplitude reaches the desired value it is necessary to reduce the gain to *exactly* 1. This can be achieved in one of two ways: Firstly, the amplifier can be allowed to go into a non-linear region where the gain decreases as the amplitude increases. Secondly, the amplifier loop gain can be dynamically adjusted to balance the signal at the desired amplitude value.

Both methods are used in practice, but if you want to make a harmonically pure oscillator then the second method is preferable. In any case you apply the *Barkhausen criterion*; the loop-gain for any stable parameter must be exactly one. The power gain is 1, the voltage gain is 1, and the current gain is 1. [The loop phase shift is exactly $2{\cdot}n{\cdot}\pi$ radians.] This must be so or the amplitude of the oscillation will change.

It is considerably easier to allow the amplifier to go non-linear as a means of defining the amplitude of oscillation. This does not necessarily mean that the output is grossly distorted, however, as this next simulation circuit shows.

FIGURE 14.7G:



When the signal gets big enough, it starts being clipped by the diodes, D1 & D2. Notice that the zeners are biassed to their specified operating conditions so that their voltages are well defined. This means that the clamp diodes are reverse biassed for most of the sinusoidal cycle. This circuit introduces a better way of simulating an oscillator. C1//R1 form a transient source which is set with an initial DC condition. This gives the simulation a 'kick' when it starts. The noise generator used for the previous simulation is not a common SPICE feature and it uses a fair amount of computational power. The 'kicking' circuit gets the oscillation up to the correct amplitude much more quickly.

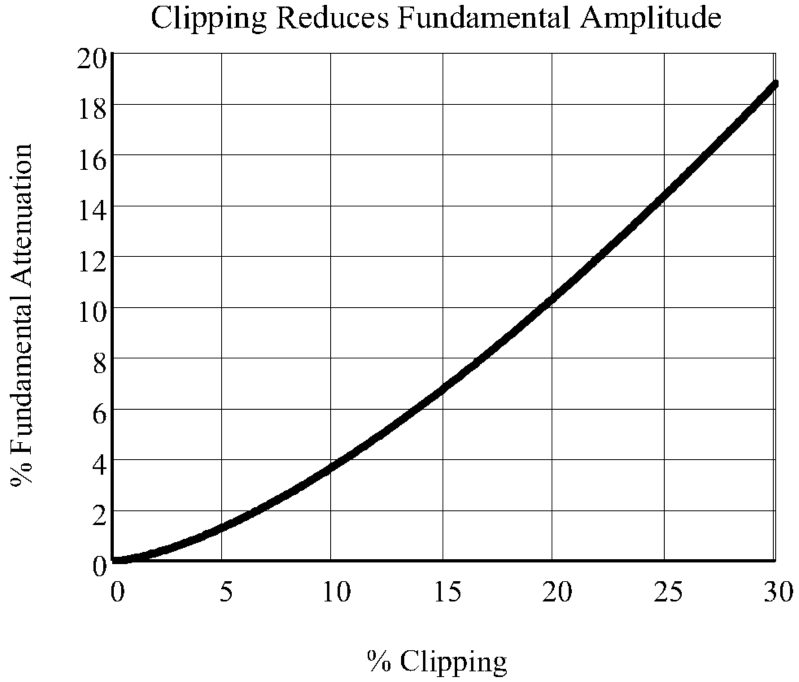
The important point is that in order for a real-world oscillator to self-start, it has to have inherent noise. This gets amplified on each pass through the loop to produce the oscillation. Since all real-world circuits have noise, an oscillator which doesn't self-start has too little loop-gain at 0° loop phase shift.

Whilst it is true that the signal at the clamp circuit is harmonically distorted, this signal then goes through a three-pole filter, attenuating the harmonics nicely. By the time the signal gets around to the output of E3, the distortion is not visible. In this simulation the third harmonic distortion is down at roughly –60 dBc. There is no second harmonic distortion in this simulation is because the zener diodes are perfectly matched.

For a practical circuit the gain might be confined to one stage to reduce its variation. If, for example, there were three stages of gain, each with its own tolerance, the overall gain would spread more. The lowest nominal extra gain for the loop would have to take into account the toleranced lowest gain that could occur in the rest of the loop. Thus the lowest [toleranced] linear loop-gain might be set at say 1.002 and therefore the maximum [toleranced] linear loop-gain might be 1.03. This would therefore give significantly more distortion. [The simulation gives –54 dBc at the third harmonic using E4 as 1.03.]

Driving a sinusoid into a clipping circuit reduces the input-to-output gain of the fundamental frequency. Notice that the third harmonic must also have a loop-gain of unity, otherwise the waveform will not be stable. The third harmonic is the most important because the worst that the clipping circuit can do is to produce harmonics of equal amplitude. The three-pole filter will attenuate higher harmonics to a larger degree. Therefore you can be assured that the biggest distortion component will be due to the third harmonic. [Even-harmonics have been neglected because they are a tolerance issue, rather than a fundamental problem with the design.]

FIGURE 14.7H:



Clipping a sinusoid changes the loop-gain of the fundamental as shown in this graph

Using more excess loop-gain gives a larger amount of distortion at the clipping circuit and therefore more distortion at the final output. However, greater filtering can be achieved by using more poles in the circuit.

You could for example use a four-pole phase shift network, each with a phase shift of 45°. Better still you could use a six-pole filter with a phase shift of 60° each and keep the phase shift network as non-inverting. However, the better solution is to electronically control the gain to the correct value.

OPAMP style variable gain amplifiers are readily available and these make earlier methods look crude by comparison. For example, it was common practice to increase the emitter current in a single transistor amplifier to increase the AC gain of the stage. This was only useful for small-signal applications, but gave very inexpensive automatic gain control (AGC) stages. It may be that a simple transistor, FET, or PIN diode variable gain stage is adequate to produce the desired result. [Stabilisation by incandescent light bulbs, or self-heating thermistors, used as non-linear elements to reduce the gain at high signal levels, should be regarded as of historic interest only, not least of which is because the settling time is so slow.]

Changing the loop-gain changes the start-up time and the amount of clipping needed to reduce the loop-gain back to unity. Provided that the clipping circuit does not introduce phase shift, the frequency will be unchanged by the loop-gain changes. The loop phase shift is another matter. It is *critical* that the loop phase shift be 360° ($2\cdot\pi$ radians) or some integer multiple thereof. Otherwise the oscillation will not be able to build up in the first place.

In any oscillator, the circuitry can be split into two distinct parts: There is an amplification device, or devices, and there is a passive phase shift network. In many texts the phase shift network is assumed to be a resonant circuit such as an LC network, a crystal, a transmission line or a cavity. But the first oscillator I have introduced is a series of RC lowpass filters; this clearly does not have a Q. A more universal measure of a stable oscillator is desirable.

The key is phase shift: the loop phase shift must be exactly $2\cdot n\cdot\pi$ radians. Therefore the phase shift introduced by the amplifier and the passive network can be equated. If the amplifier phase shift changes by $+\delta\phi$ then the network has to introduce a $-\delta\phi$ phase shift to compensate. This is done by changing the frequency: $\Delta f = \delta\phi \cdot \frac{df}{d\phi}$

In normalised form this is $\frac{\Delta f}{f_0} = \delta\phi \cdot \frac{1}{f_0} \cdot \frac{df}{d\phi}$

This gives a figure of merit for the oscillator network as $\boxed{M_0 = f_0 \cdot \left|\frac{d\phi}{df}\right|}$

This figure needs to be as large as possible. Assuming that the amplifier does not change its phase shift with frequency very strongly, compared to the network, then there are two contributions to $\delta\phi$: the change from the network and the change from the amplifier. These can occur due to temperature change or component drift. The result is

$$\frac{\Delta f}{f_0} = \frac{\delta\phi_{AMP}}{M_0} + \frac{\delta\phi_{NET}}{M_0}$$

If this is a low frequency oscillator, using a really fast amplifier, you could suppose that the phase drift due to the amplifier is negligible compared to the phase drift of the network. In this case only the *self-stability* term $\frac{\delta\phi_{NET}}{M_0}$ is left.

At the very highest frequencies it may be that the phase drift due to the amplifier is much larger than the phase drift due to the network. In this case the frequency stability is improved by individually improving the phase drift of the amplifier and the figure of merit of the network.

A single-pole filter has an AC transfer function of: $T = \frac{1}{1 + j \cdot \frac{f}{B}}$

The corresponding phase shift (magnitude) is then given as $\phi = \arctan\left(\frac{f}{B}\right)$

Giving $\frac{d\phi}{df} = \frac{\cos^2(\phi)}{B}$

With N identical buffered stages giving a total of 180° phase shift [π radians], $\phi=\pi/N$, giving the figure of merit as:

$$M_0 = f_0 \cdot \left|\frac{d\phi}{df}\right| = N \cdot \frac{f_0}{B} \cdot \cos^2\left(\frac{\pi}{N}\right) = N \cdot \tan\left(\frac{\pi}{N}\right) \cdot \cos^2\left(\frac{\pi}{N}\right) = \frac{N}{2} \cdot \sin\left(\frac{2\pi}{N}\right)$$

Each stage can be represented by the transfer function $T = \frac{1}{1 + j\Omega}$. The phase shift per stage being $\phi = \arctan(\Omega)$ and the attenuation loss being $A = \sqrt{1+\Omega^2}$.

Thus $\tan(\phi) = \Omega$ and $A = \sqrt{1 + \tan^2(\phi)} = \frac{1}{\cos(\phi)}$.

The overall gain required is $A^N = \frac{1}{\cos^N\left(\frac{\pi}{N}\right)}$

FIGURE 14.7i:

This simple LC oscillator simulates at 5 MHz. The open-loop phase shift is fairly symmetric about the resonant point because the Q is quite high (≈30). From the $d\phi/df$ plot, if the amplifier phase shift is < ±40°, $d\phi/df$ is relatively constant.

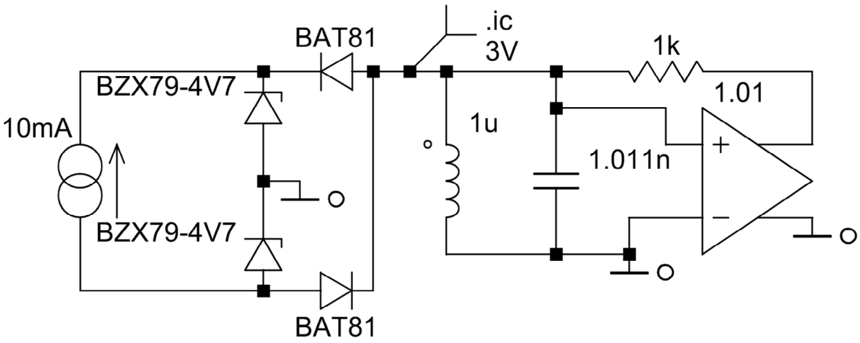
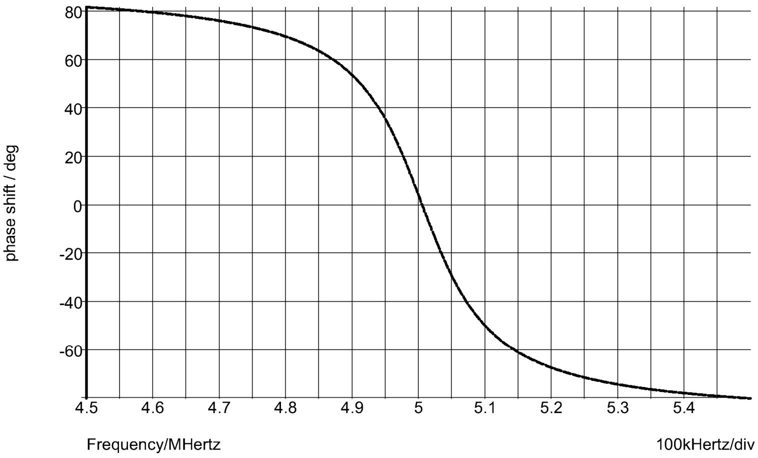


FIGURE 14.7J:



The figure of merit for this resonant circuit can be estimated by scaling off of the graph. The centre frequency is 5 MHz and the phase shift is roughly 40° for 60 kHz. Hence M0 is approximately 60. That is more than 10× better than the best of the previous RC networks.

$$\phi = \arctan\left(Q \cdot \left[\frac{f_0}{f} - \frac{f}{f_0}\right]\right) \quad \text{giving} \quad \frac{d\phi}{df} = \frac{Q \cdot \left[-\frac{f_0}{f^2} - \frac{1}{f_0}\right]}{1 + Q^2 \cdot \left[\frac{f_0}{f} - \frac{f}{f_0}\right]^2}$$

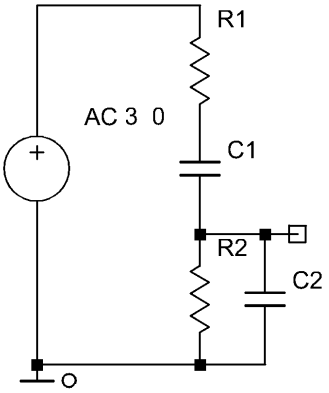
$$M_0 = f_0 \cdot \left|\frac{d\phi}{df}\right| = \frac{Q \cdot \left(1 + \frac{f_0^2}{f^2}\right)}{1 + Q^2 \cdot \left(\frac{f_0}{f} - \frac{f}{f_0}\right)^2}$$

Evaluate this function when $f = f_0$

$$\boxed{M_0 = f_0 \cdot \left|\frac{d\phi}{df}\right| = 2Q}$$

Even the poorest of LC networks [Q > 2] is therefore better than the best RC filter chain.

FIGURE 14.7K:



This is the 'working' arm of a *Wien bridge* and when used with a single-ended amplifier forms a [half-] Wien bridge oscillator. Analysis shows that the optimum figure of merit for this network is 1. This is a lot worse than the optimum figure of merit for cascaded buffered single-pole filter sections (figure of merit π). If R1=R2 and C1=C2 the figure of merit for the Wien bridge is only 0.667. This is twice as bad as the buffered 3-pole phase shift network and therefore the Wien bridge is not as good as its reputation might suggest.

Summary of Simple Oscillator Types

TYPE	$M_0 = f_0 \cdot \left\| \frac{d\phi}{df} \right\|$	Required Voltage Gain
3-stage identical unbuffered phase shift oscillator	1	29
3-stage buffered inverting phase shift oscillator	1.3	8
Half Wien bridge oscillator	0.7	3
10-stage identical unbuffered phase shift oscillator	1.6	12.5
6-stage buffered inverting phase shift oscillator	2.3	2.4
Inverting propagation delay oscillator	3.1 [= π]	1
Non-inverting propagation delay oscillator	6.3 [= 2·π]	1
Single LC resonant circuit oscillator	2×Q	1

For more complicated narrow band devices such as SAWs and crystals, the figure of merit is double the Q, as is the case for the single LC resonant circuit.

There are also simple oscillators required for microcontrollers which are often not required to be accurate. These either use inexpensive crystals or ceramic filters. They are not technically demanding applications in the sense of the quality of the oscillation. The prime requirement is that the oscillator starts and runs reliably over large production runs. The key thing here is to follow the manufacturer's recommendations and application notes. Typically you take the design and deliberately increase and decrease the component values to find the limits where the oscillator only just runs; the values are then chosen as the geometric mean of the limits.

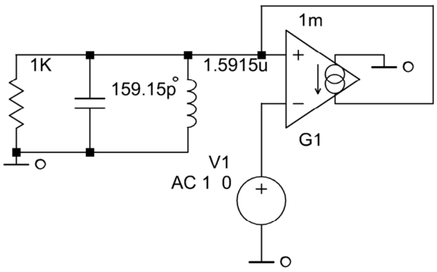
Phase Noise

An oscillation builds up by repeated amplification of noise when the phase shift around the loop is exactly $2n\cdot\pi$ radians. Since the noise in the amplifier covers a much larger range of frequency than just the oscillation frequency, what happens to it? More specifically, how is the noise gain related to the Q?

This simple simulation model helps to answer the question "how is noise gain related to Q". The resonant circuit on its own has a centre frequency of 10 MHz, a Q of 10, and therefore a bandwidth of 1 MHz.

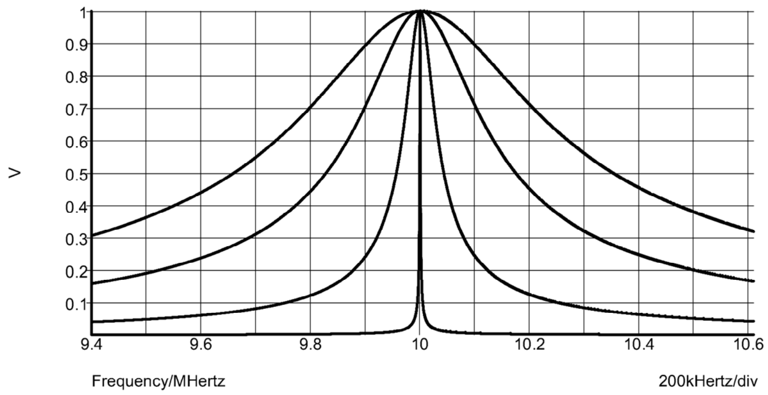
The amplifier is a voltage controlled current source, but this simulation circuit still models all LC oscillators.

FIGURE 14.7L:



At resonance, the parallel circuit input impedance is 1 kΩ, so a gain in the amplifier of 1 mA/V will give the critical loop-gain of 1. For this simulation the gain of the amplifier is slowly stepped up to near the critical gain and the resulting response viewed. For each gain setting, the voltage source V1 is adjusted to keep the resonant peak voltage constant, making the comparison easier.

FIGURE 14.7M:



According to this model, as the loop-gain approaches unity, the Q approaches infinity and the bandwidth drops to 0. This is a misleading result. There is gain of the noise close to resonance, and since the amplitude at resonance is limited, the Q becomes high, but not infinite.

FIGURE 14.7N:

This new simulation model is the Thévenin equivalent of the previous one. In this model the loop-gain factor is more readily seen as approaching unity. As before the source Vn can be swept to investigate the performance of the oscillator. Think of Vn as the equivalent noise generator of the amplifier.

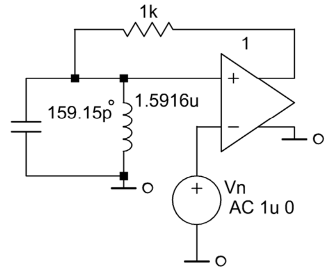
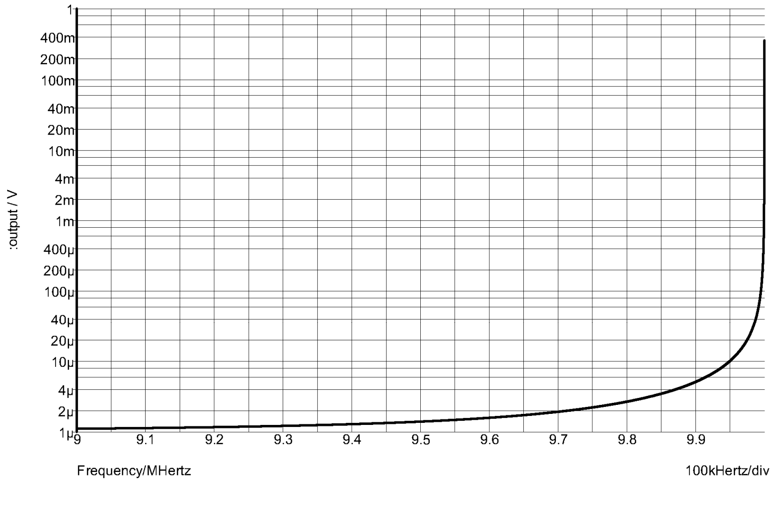


FIGURE 14.7O:



The simulated output of the oscillator circuit above shows an asymptotic approach to 1 µV as the frequency is decreased. This is the noise generator Vn coming straight out without any amplification by the loop.

The voltage at resonance, however, is not defined by this method. The loop-gain can be set exactly to unity on this small-signal analysis and at the resonant point the output amplitude is only limited by rounding errors in the simulator.

The peak amplitude of the oscillator is actually limited to some value $\hat{V}_{RMS}$ by the power rails or by clipping circuitry. The bandwidth and Q are therefore defined by the noise gain slightly off resonance.

The transfer function across the resonant network can be written down by inspection:

$$T = \frac{\left(\frac{1}{j\omega L} + j\omega C\right)^{-1}}{\left(\frac{1}{j\omega L} + j\omega C\right)^{-1} + R} = \frac{1}{1 + R\left(\frac{1}{j\omega L} + j\omega C\right)} = \frac{1}{1 - j\left(\frac{R}{\omega L} - \omega CR\right)} = \frac{1}{1 - jQ\left(\frac{f_0}{f} - \frac{f}{f_0}\right)}$$

… where f_0 is the resonant frequency and the circuit Q has been evaluated from the component values. Using V_m as the output voltage at frequency f due to the noise voltage V_n, $V_m = V_m \times T + V_n$. Rearranging this gives an explicit equation for the output voltage in terms of the noise and the network transfer function:

$$V_m = \frac{V_n}{1-T} = V_n \frac{1 - jQ\left(\frac{f_0}{f} - \frac{f}{f_0}\right)}{1 - jQ\left(\frac{f_0}{f} - \frac{f}{f_0}\right) - 1} \quad \rightarrow \quad \boxed{V_m = V_n\left(1 + j\frac{1}{Q\left(\frac{f_0}{f} - \frac{f}{f_0}\right)}\right)}$$

A high-Q network therefore gives less amplification of the noise and consequently less phase noise. Using this new equation, the limiting value of Q for the overall circuit can be determined. Defining the voltages as RMS values, and writing the maximum operating output voltage as $\hat{V}_O$, the bandwidth of the oscillator is given when the amplified noise is at the half power (3 dB) points.

Using f_L as the lower frequency 3 dB point gives $\frac{\hat{V}_O}{\sqrt{2}} = V_N \left| 1 + j\frac{1}{Q\left(\frac{f_0}{f_L} - \frac{f_L}{f_0}\right)} \right|$

$$\frac{1}{2}\left(\frac{\hat{V}_O}{V_N}\right)^2 = 1 + \frac{1}{Q^2\left(\frac{f_0}{f_L} - \frac{f_L}{f_0}\right)^2} \quad \text{giving} \quad \left(\frac{f_0}{f_L} - \frac{f_L}{f_0}\right) = \frac{1}{Q\sqrt{\frac{1}{2}\left(\frac{\hat{V}_0}{V_N}\right)^2 - 1}}$$

Since the output voltage will always be much larger than 32× than the noise, the unity term within the square root sign can be neglected without any significant loss of accuracy (<0.1% error).

$$\left(\frac{f_0}{f_L}-\frac{f_L}{f_0}\right)=\frac{V_N\sqrt{2}}{Q\cdot\hat{V}_0}$$

This equation is of the form $x-\frac{1}{x}=\Delta$,

which can be put into standard quadratic form $x^2-x\Delta-1=0$,

the solution of which is $x=\frac{\Delta+\sqrt{\Delta^2+4}}{2}=\frac{\Delta}{2}+\sqrt{1+\frac{\Delta^2}{4}}\approx 1+\frac{\Delta}{2}+\frac{\Delta^2}{8}$.

Because the response is symmetrical about the resonant frequency, the bandwidth is $2(f_0-f_L)$ and the effective Q is therefore

$$Q_E=\frac{f_0}{2(f_0-f_L)}=\frac{1}{2\left(1-\frac{f_L}{f_0}\right)}.$$

Now $\frac{f_L}{f_0}=\frac{1}{x}=\frac{1}{1+\frac{\Delta}{2}+\frac{\Delta^2}{8}}\approx 1-\frac{\Delta}{2}$ giving $\boxed{Q_E=\frac{1}{\Delta}=\frac{Q}{\sqrt{2}}\cdot\frac{\hat{V}_O}{V_N}}$

An oscillator therefore massively increases the overall circuit Q.

Phase noise is expressed in units of dBc/Hz, a slightly tricky unit. It should really be dBc in a 1 Hz bandwidth, since you cannot divide the number of dB by the number of Hz, but dBc/Hz is the standard notation.

offset	phase noise
1 Hz	–112 dBc/Hz
10 Hz	–142 dBc/Hz
100 Hz	–160 dBc/Hz
1 kHz	–167 dBc/Hz
10 kHz	–170 dBc/Hz
100 kHz	–170 dBc/Hz

These typical phase noise figures are from a Spectra Dynamics Inc LNFR-100 ultra-low noise frequency reference oscillator at 10 MHz. There are not enough points to determine the shape of the response (see graph overleaf), but it is clear that the slope is not constant. In Leeson's model,[4] the phase noise in an oscillator is given by three *asymptotic* slopes: far from the open-loop bandwidth of the resonant network the phase noise is constant with frequency. Within the open-loop bandwidth of the resonator, the phase noise increases at a rate of 6 dB/octave (20 dB/decade). Below the flicker noise frequency of the amplifying device, the phase noise increases at a rate of 9 dB/octave (30 dB/decade).

Rather than using the absolute frequency, phase noise is always presented in terms of the offset frequency from the carrier. Denoting the offset frequency by f_m, the noise transfer function becomes:

$$V_m=V_n\left(1+j\frac{1}{Q}\left(\frac{f_0}{f_0-f_m}-\frac{f_0-f_m}{f_0}\right)^{-1}\right)=V_n\left(1+j\frac{1}{Q}\left(\frac{1}{1-\frac{f_m}{f_0}}-1+\frac{f_m}{f_0}\right)^{-1}\right)\approx V_n\left(1+j\frac{f_0}{2Qf_m}\right)$$

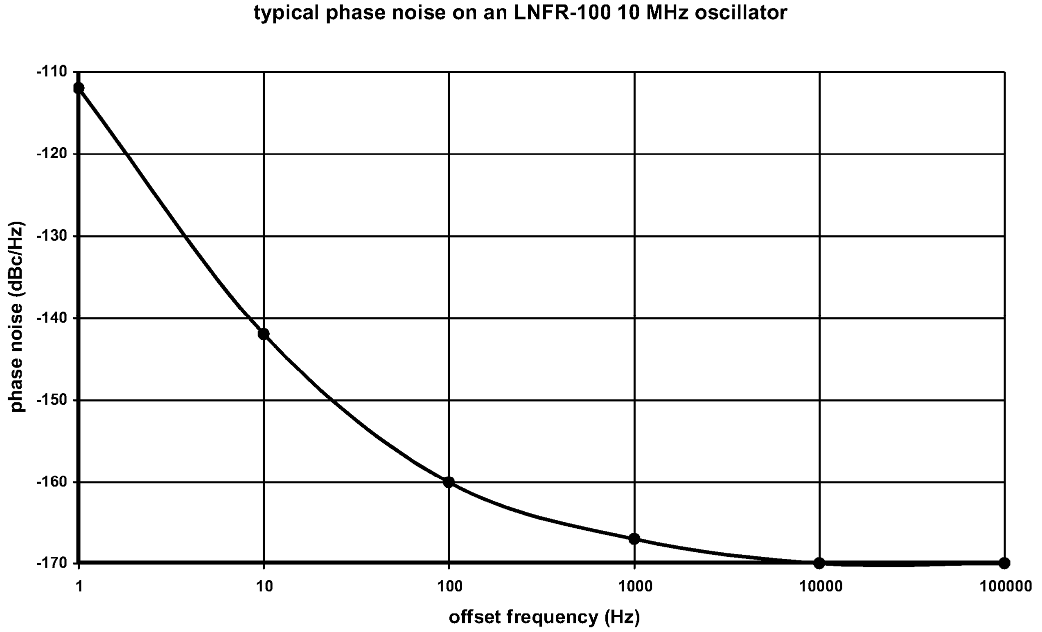
[4] D.B. Leeson, 'A Simple Model of Feedback Oscillator Noise Spectrum', in *Proceedings of the IEEE: Letters*, 54, no. 2 (Feb 1966), pp. 329-330.

This approximation is only good (<0.5% error) when Q > 50. It is this approximate form that is given in Leeson's paper, although the phase noise is given in terms of the power spectrum, this being defined as the voltage squared divided by 1 Ω.

When a reference oscillator is multiplied up in frequency using a phase-locked loop (***PLL***), the phase noise of the resultant oscillator is at best multiplied by $20 \cdot \log_{10}(N)$, where N is the multiplication factor. Multiplying the 10 MHz output to 100 MHz will therefore give rise to at least an additional 20 dB of phase noise. In practice another 3 dB or more will be contributed by the phase comparator and divider in the PLL.

If this 10 MHz source is now multiplied up to 10 GHz the phase noise will be increased by 60 dB just due to the multiplication. At 10 GHz a *dielectric resonator oscillator* (DRO) can be used to make a stable oscillator. Any 10 GHz oscillator with phase noise worse than −100 dBc/Hz @ 10 kHz offset could be replaced by a multiplied crystal source. However, at least one manufacturer claims −135 dBc/Hz @ 10 kHz offset, which is considerably better than a multiplied crystal source.

FIGURE 14.7P:



14.8 Transistor-Level Design

By "transistor-level design" I mean using discrete transistors to make amplifiers, signal conditioners &c. The key thing to say about this is **don't do it**! Whilst it used to be the case that adding transistor input or output stages around opamps could improve their performance, since just before the turn of the century modern opamp performance has become so good that such hybrid amplifier designs are no longer either necessary or commercially viable.

You may still need to do designs using discrete transistors for oscillators or microwave/mm-wave amplifiers, but that is a very specialist area. Be aware that somebody who has done discrete transistor design before will be at least 10× faster at it than a novice. In fact this speed factor can be as bad as 100× or even tending to infinity. As the spec gets harder to meet, a novice may **never** get a complicated design to work. This is a hard truth, but much of the expertise is just not written down. Sometimes the only solution is to call for help. Just don't leave it so late that the consultant doesn't have a chance to fix the problem before the deadline. We all like to think that we can solve any problem, and do any job. The reality is that most of us probably could, if given enough time and resources. The trouble is that the company might have gone bankrupt waiting for us to get there!

Some key problems to look out for in discrete transistor design are:

- Imperfections in the system response which change slowly at rates of anything from tens of seconds to tens of microseconds, these drifts being due to thermal time constants in the devices or across the PCB. *These effects are not modelled at all by standard SPICE based simulators*.
- Spurious oscillations at frequencies up to half the f_t of the active devices.
- Latch-up conditions, where too large an input signal causes the circuit to get stuck into an incorrect state.
- Start-up problems, where the power rail rise-times give unexpected results.
- Interactions between components. You test function *A*, it is ok. Function *B* is faulty, so you change something. Function *B* now works. You *must* recheck function *A*, because it may have been made faulty in the process of fixing *B*.

Obviously I can't think of all the weird things that discrete circuitry can do. You really do have to poke about with the circuit and see if it is doing what you wanted, or something 'extra'.

You will still need to use simple transistor circuits to perform simple functions. Digital engineers call the function "glue logic" when they use the odd inverter, AND gate or whatever amongst all the big gate arrays and field programmable logic arrays in the design. Well 'analog glue circuitry' is also used. You can use transistors to switch relays, to selectively remove power from parts of a circuit, to disable oscillators and other noisy circuitry in certain quiet operational modes, to reduce power when some circuitry is not needed &c. The point is that these are not demanding analog applications. The transistors are being used as on/off switches. In these applications the discrete component solution will be very much cheaper than an integrated solution. It will also perform better.

CH15: measurement equipment

15.1 The Moving Coil Meter

The first meter used to measure electric current using the magnetic effect of a current was devised by Ampère in 1820; he coined the name *galvanometer* for it. Nowadays we would call any current measuring instrument an *ammeter*, in honour of Ampère, although the name galvanometer is also still used, especially for sensitive fixed range ammeters.

Ampère's galvanometer consisted of a magnetic compass placed near to the current-carrying conductor. The idea was to note the deviation of the needle when the current was applied and this would be a measure of both the direction and the amplitude of the current. Schweigger's *multiplier*, also of 1820, used a 100 turn coil of wire around the compass needle to get enhanced sensitivity.

The moving coil meter is a nice, simple piece of equipment, often referred to as an 'analog meter' by comparison with a digital multi-meter (DMM). Note that inexpensive DMMs have only been available since around 1970.

A moving coil meter has several features that still make it superior to its more modern digital equivalent. Firstly, a moving coil meter is not a source of noise. It just sits there quietly doing nothing, electrical speaking. A digital meter, on the other hand, is sampling, pulsing, converting and displaying; these processes will push small noise current spikes back out of the input terminals. These currents may or may not be significant for the measurement you are making.

Comparing a moving coil meter to a mains-powered DMM, notice that there is excellent isolation from the moving coil meter to earth {ground; mains circuit protective conductor}. Whilst it is usual for mains powered DMMs to have several hundred volts of isolation, a moving coil meter standing on a big block of polystyrene {insulator} can measure currents that are tens of thousands of volts above earth potential. (You can also do this same trick with a battery powered hand-held DMM.)

When measuring resistance, the moving coil meter will pump out a steady direct current. A DMM, especially a hand-held, may well use a pulsed current, either due to the sampling technique used, or to save on battery power. In any case, even expensive DMMs can give spurious measurements when used to measure resistances in highly inductive or highly capacitive circuits. Measuring the DC resistance of a transformer winding, for example, is actually a difficult task for a DMM; it has been known for the current source in the linear ohms system to become unstable with such a heavy inductive load.

Moving coil meters are not obsolete; they still have a place in a modern lab. When measuring voltage or current they have three great virtues

- No mains socket required on bench.
- Batteries don't go flat if left on permanently
- Doesn't generate any RFI or noise.

Given that the needle on a moving coil meter is not in contact with the scale, the observer's position will affect the exact reading taken, the technical name for which is a *parallax error*. There are two ways to minimise this error; the *mirror scale* and the deep needle.

With a mirror scale, you move your head until the needle's reflection gets hidden by the needle itself. This guarantees that you are looking at right angles to the scale. With the deep needle approach, which is cheaper but less effective, you move your head to the position where the needle seems to be the thinnest. Again you are perpendicular to the scale.

Note that a *mirror galvanometer* is not the same thing as a mirror scale galvanometer. Rather than use a needle, a mirror galvanometer uses a small mirror to reflect a beam of light onto a scale. This technique can produce a much more sensitive meter, not least of which is because the light beam can be projected a distance of more than a metre.

EX 15.1.1: A moving coil meter has 20 kΩ/V written on the front face. It is reading 37 V on the 100 V range. What current is it drawing from the circuit being measured?

EX 15.1.2: You measure a circuit with an output impedance of 12.4 kΩ using a 20 kΩ/V moving coil meter set to the 10 V range. What measurement error do you get from the loading effect alone?

EX 15.1.3**: A moving coil ammeter has a full scale ***burden voltage of 500 mV on its DC current ranges. The meter has no over-range capability. [Full Scale= Full Range in this case.]

a) What is the volt drop across the meter when measuring 100 mA on the 1 A range?
b) What is the volt drop across the meter when measuring 100 mA on the 100 mA range?
c) What is the resistance of the meter on the 100 mA range?

EX 15.1.4: A moving coil ammeter is placed in series with a logic rail power supply. Evaluate the following statement carefully: “The ammeter reads the wrong current because of its burden voltage.”

To illustrate how sensitivities have improved over the years, here is a table representing the sensitivities of the moving coil meters made by Avo Ltd.[1]

Model	Approx. year	Full Scale current	DC Sensitivity
Avometer	1923	12 mA	83 Ω/V
DC Avometer	1927	6 mA	167 Ω/V
36 Range Universal	1935	3 mA	333 Ω/V
Avometer Model 7	1936	1 mA	1 kΩ/V
HRM and HR2	1946	50 μA	20 kΩ/V
Avometer 8 Mk I	1950	50 μA	20 kΩ/V

The sudden change in sensitivity in 1946 was due to the availability of new magnetic materials such as ALNICO.

[1] R.P. Hawes, *History of the Avometer*

Whilst the DC sensitivity is 20 kΩ/V, the AC sensitivity, even on a modern moving coil meter, can be very low. For example:[2]

3 V range	100 Ω/V	
10 V range	1000 Ω/V	
30 V range	2000 Ω/V	(higher ranges are the same)

100 kΩ/V moving coil meters are readily available, but there is no impetus to improve this figure because high accuracy measurement (better than 0.1%) is exclusively the domain of digital meters.

HISTORICAL NOTE: Moving coil meters with higher resolutions and accuracies have been BIG. This made it easier to read the needle position. A 0.5 m scale was not unusual. To get an even larger scale, a mirror (projection) scheme was sometimes used. A mirror was attached to the meter movement and a light beam bounced off the mirror onto a large cylindrical scale some distance away. Such a meter could have a 3 m scale length, giving much greater measurement resolution. The user was essentially *inside* the meter, as the scale was part of the (darkened) room!

15.2 The Digital Meter

The first thing to say about using a precision digital meter, or any other complicated piece of equipment, is to find the manufacturer's manual and read it. The manual may not be easy to find. It may have been hidden away for 'safe keeping' in another part of the building or just lost. Others will tell you that you don't need it and that all you have to do is this, this and this.

Only somebody who had themselves never read the manual would suggest this! There can be a wealth of information in the user's manual. It was written so that you can get the most out of that particular device. Ignore it at your peril. Don't try to read the whole manual from cover to cover. Skim across it, looking for the useful information.

A DMM is a Digital Multi-Meter; a DVM is a Digital Volt-Meter. Manufacturers of expensive DMMs sometimes call them "multifunction DVMs" to show that they are different to cheap hand-held meters!

Let's not worry too much about 3½/4½ digit meters. [See DMM scale sizes in the Glossary] Let's talk about the 5½/6½/7½/8½ digit meters. If you are using these then you need to know what you are doing in order to get accurate readings.

Rather than listing all the things that you need to know about your DMM, let me instead take the points one at a time. Some of these spec points may be labelled on the equipment, but many more will not be. You are getting "insider information" here because I used to design 7½ digit DMMs.

It is quite common for long scale DMMs to have extremely high input impedances on the ranges around 10 V and below. This gives better accuracy by reducing the loading effect on the source. On the higher ranges, it is no longer possible to maintain the high input impedance because a resistive input attenuator is used. The input will then drop to something like 10 MΩ. This can be illustrated with an exercise.

[2] AVOMETER Model 8, Mk 7; AVO international.

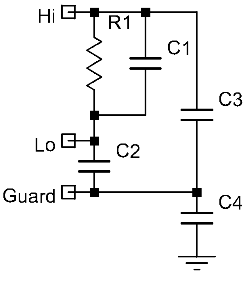
***EX 15.2.1:** A precision DVM has an input spec on its lower ranges (<12 V) of bias current <50 pA, input resistance >10 GΩ.

a) The meter is zeroed with a short-circuit, and then a shielded 10 MΩ resistor is connected across the terminals in place of the short-circuit. The meter has been given adequate time to stabilise and all other conditions are as stated in the manufacturer's environmental conditions. The meter reads +0.3 mV. Is the DMM in spec?

b) Another meter of the same type is connected to a quiet low-impedance (<1 mΩ) voltage source through a 10 MΩ resistor. The voltage source is set at +10 V. This is well within the lower ranges of the DMM. Give limits for the current in the resistor.

DMM input terminals are often called *High* and *Low*, abbreviated to Hi (red) and Lo (black) respectively. Do not think that the terminals of the DMM are the same and can therefore be connected either way around. They are not the same, and should never be treated the same. It is usual to have a metal *guard box* inside the DMM; this is either connected to a separate guard terminal or is connected to the Lo terminal internally.

FIGURE 15.2A:



Establish the values of the capacitors from the DMM spec, from direct measurements, or by enquiry of the manufacturer. Rather than have capacitances to ground {earth}, the DMM has capacitance to the internal guard box. Typically C2 >> C3.

By connecting the guard box to the common-mode source, measurement errors can be reduced. It is important to realise that there will be a definite AC current in the guard wire.

Direct measurement of the capacitances should be done at ≤1 kHz since a DMM is a low frequency instrument., probably having a bandwidth below a few tens of hertz. Using a 100 kHz test signal to measure the input capacitance may result in a misleading result due to the internal amplifiers slew rate limiting.

Since the instrument will be powered from the mains supply, the mains transformer will probably have a grounded {earthed} screen between the primary and secondary winding. Additionally there will be a guard screen between the ground screen and the secondary windings.

Measured values from one example of a 5½ digit DMM on its 1000 V range were: guard to earth= 700 pF, Lo to guard= 1.1 nF, Hi to guard= 150 pF, Hi to Lo= 130 pF. If R1 is >10 GΩ on the lower voltage ranges, expect C1, C2 and C3 to also change when the input attenuator is removed.

Since the DMM input capacitance may be 100 pF, with the leads adding at least as much, this load capacitance may make the circuit under test unstable {oscillate}. In this case, a resistor of between 100 Ω and 10 kΩ is needed between the measured point and the test leads to stop the oscillation.

The leads used to make measurements can be very important when measurement resolution below 300 μV is required. Individual pieces of ordinary PVC covered stranded tinned copper wire are not suitable for precision measurements. Firstly, it is necessary to get the terminals of the DMM and the wire at the same temperature in order to minimise ***thermal EMFs***. This is best accomplished by using gold plated spade connectors. These give a reliable, temperature stable and repeatable interconnect.

Secondly, the lead wires should either be a twisted pair or inside a screened cable; connect the screen to the DMM guard terminal, or to the Lo input if there is no explicit guard. The twisted pair will help to reject {minimise} stray magnetic pickup and the screened cable will help to reject electric fields.

Another useful technique is to shield the connection terminals from drafts. Changes of temperature will produce noise, depending on how badly matched the materials are to each other. Even a gold-copper interface will give 0.3 µV/°C, producing an unacceptable amount of noise in the sub-µV region when subjected to drafts. Big heavy terminals help to smooth out any rapid fluctuations.

You can make a measurement with a long scale DMM and the reading can be virtually noise free, and yet when looking at the same signal on a scope, lots of AC ***mains*** frequency noise can be seen. The reason is that long scale DMMs deliberately integrate over one or more power line cycles in order to reject this problem frequency. The DMM display will therefore not appear noisy if the noise is at the power line frequency.

For DMMs that measure true RMS, the “RMS converter” takes an incoming signal and makes an equivalent DC value out of it. It is this DC value which is measured. If the RMS converter is AC coupled then you would expect a lower frequency operational limit. However, if the RMS converter was able to work DC coupled you might reasonably expect that the RMS readings would be valid down to any arbitrarily low frequency. This idea is false. There is an additional low frequency limit below which the reading is not stable and is no longer an RMS value. In the limit, on an AC waveform, the DMM would just be reading sampled values of the ‘full-wave rectified’ version of the waveform. Averaging the readings yourself gives half-cycle mean, *not* RMS. This effect happens with both thermal converters and the log-antilog types. Check the manual for the low frequency limit.

All metering devices that do not display the input signal, but only give an answer as a number on a scale can give misleading answers. This example demonstrates the problem.

EX 15.2.2: An accurate DMM is used to measure a test point on a circuit. The test point is at a DC level of 100 mV with 1 V ptp of 50 Hz signal, the local mains frequency. The DMM is fixed on its 10 V range. The DMM specifies line frequency rejection as better than 100 dB. What reading does the DMM display? (Neglect measurement uncertainties and tolerances for this question).

There is another problem with integrating DMMs and that is internal clipping {limiting}. If you put a wildly varying DC signal into a DMM and it happens to overload the internal amplifiers at certain points during the cycle, there may be no evidence of this on the display. The meter will no longer be reading the mean voltage and will be in error by some unknown amount.

To measure the mean value of a rectangular waveform, for example, you should ideally filter the signal external to the DMM, thereby reducing slew limit and clipping problems inside the DMM. In any event, you should ensure that the peak signal going into the DMM on DC does not exceed the full scale value. As a test you could change to the next range up and see if the reading changes by several percent. This is a clear indicator of a measurement problem.

An even more common problem with DMMs is applying an unsuitable waveform to a true RMS converter. Again it relates to clipping in the internal amplifiers. True RMS converters have a spec for this, known as ***crest factor***, the ratio $\frac{peak\ signal}{RMS\ signal}$. Crest Factors can be expressed as ratios such as 3:1, or as just 3. Typical limits are from anywhere from 3 to 7 at full range, with 7 being the better spec. If the crest factor spec is 5 on the 1 V range, this means that you must not put in a 1 V RMS signal whose peak value exceeds 5 V. The DMM will have a better crest factor handling capability at a lower part of the scale, as the amplifier stages will still be clipping at the 5 V level. However, if you insist on putting in a 100 mV RMS signal with peaks up to 5 V, a crest factor of 50, the DMM accuracy will no longer be defined.

Unless you are using a *thermal transfer standard*, you must accept that a true RMS meter has an unspecified uncertainty on the measurement of signals whose crest factor exceeds the instrument rating. The reading may be consistent and repeatable (on that DMM), but it will not be ***traceable*** to national standards. The only way around this problem is to individually calibrate the waveform in question against a genuine thermal transfer standard to restore the traceability.

A typical set of questions concerning the DMM you will be using:

- What is the input resistance of the DMM on DC volts?
- If the input resistance is very high, what is the bias current?
- Does the input resistance change on the higher DC ranges?
- What is the input capacitance of the DMM?
- What is the capacitance from the guard terminal to ground {earth}?
- What is the frequency response of the AC Voltage measurement?
- Is the AC measurement true RMS or RMS-calibrated half-cycle mean values?
- What is the AC coupling high-pass corner?
- Can the AC converter handle DC coupled DC + AC signals?
- What is the low frequency limit for DC + AC signals?
- What is the crest factor on full range AC signals?
- What is the integration time for the acquisition system?
- Is there a local guard facility or is the guard connected to the Lo terminal?
- Is there an Ohms guard facility and how is it used?
- How do you switch from 2-wire to 4-wire ohms measurements?
- What is the measuring current on the ohms ranges and is it DC or pulsed?

For your immediate application you will not need to know each and every one of the answers to these questions. Generally you would just use the manual to find the answer to your specific question at that time. All I am trying to tell you is that these questions often have to be considered if you are to make sensible, useful and accurate measurements. They are also useful questions to ask when buying an expensive DMM.

You will be trying to make an accurate (traceable) measurement with your DMM. There is no *correct* answer for this reading available anywhere. The only way you can be assured that the reading is correct is to ensure that the reading has been taken in a correct manner, taking into account system-induced errors. In general, any measurement has an associated uncertainty {range of possible error}. A good measurement technique will minimise this uncertainty.

15.3 The Oscilloscope

As with the DMM, you should first find the manual and skim over it to find the useful sections. In reading it you will also discover what makes a manual good. Stating that pressing the “filter” button turns on the filter is not helpful! Why is turning on the filter worthwhile? With a filter the use is probably obvious, but specialist equipment has specialist functions whose purpose is non-obvious; as the designer you should explain why this particular function is useful.

Physics books love to call scopes CROs, *cathode ray oscilloscopes*. The student is expected to know that a *cathode ray* is an early name for an electron beam. In this book I use the abbreviated term *scope*, not least of which is because many modern scopes use liquid crystal displays (LCDs), not electrostatic deflection cathode ray tubes.

There are three distinct types of scopes:

- real-time (analog) scopes
- storage scopes
- sampling scopes

This equation gives the (complex) normalised frequency-domain transfer function, *T*, of any circuit with a single-pole low-pass bandwidth *B*. *T* is the normalised ratio of output voltage (or displayed voltage) over input voltage. The normalisation process makes the low frequency transfer function unity.

$$\boxed{T = \frac{1}{1 + j\dfrac{f}{B}}}$$

***EX 15.3.1:** A scope has a spec of ±3% and a bandwidth of 20 MHz.

a) What risetime would you expect to measure if you applied a correctly terminated 1 MHz square wave, having 200 ps risetime, to the scope input?

b) If you measured the amplitude of a 10 MHz (correctly terminated) sine wave what error would you get?

For safety reasons the inputs of scopes are normally solidly bonded to the *circuit protective conductor* {CPC; earth; ground}. If the BNC outer is grounded, you cannot connect it to power rails within equipment under test, unless that power rail is itself grounded, or the whole system is floating relative to ground.

The key thing to find out about your scope is whether or not the BNCs are grounded, since a small proportion of modern scopes have isolated inputs. It is also a good idea to test the ground bond with a meter. Ground leads can fall off, or be deliberately taken off, and this will make the equipment dangerous. In any case it is essential that any workbench is also protected by some form of residual current device {RCD, *earth leakage circuit breaker*; also known as a *ground fault circuit interrupter*, GFCI}. If any mains current passes through a ground {earth} path, rather than back through the supply wires, the breaker trips, saving you from possible electrocution.

For all scopes, especially those with bandwidths in excess of 200 MHz, the bandwidth is not necessarily constant with vertical sensitivity. The bandwidth often gets reduced as the sensitivity is increased. In one specific example of an expensive scope[†] the specified bandwidth is 1 GHz on the 10 mV/div range, but drops to 500 MHz on the 1 mV/div range.

Another common problem with all types of scope is *attenuator compensation*. The input of the scope can only take a certain amount of signal directly. Depending on the model of scope the "straight through" ranges will be 50 mV/div and below for higher bandwidth scopes and perhaps 200 mV/div and below for lower bandwidth scopes. Unfortunately this information will not be given in the instruction manual.

You may hear the attenuator relays click as you change up to the first attenuated input range or you may spot that the noise seems to get worse as you change from say the 50 mV/div range to the 100 mV/div range. A high impedance (1 MΩ) input attenuator has been switched in. Now this attenuator consists of precision resistors and adjustable capacitors. If the resistors drift by 0.5%, and the capacitors remain stable, the pulse response will undershoot or overshoot by this amount. On an older scope which has not has its input attenuators re-adjusted, the pulse response on the higher ranges could easily be in error by a percent or more.

Another source of error is input capacitance. If the capacitance of the input attenuators is not matched to the unattenuated input capacitance, the scope will give pulse response errors when used with a 10:1 passive probe. Scopes which have been "calibrated" at a discount price may therefore not be correctly adjusted, and this is something to watch out for.

Real-Time Scopes:

A "real-time" scope is the original analog version having a Cathode Ray Tube (CRT). Light is produced by an electron beam striking a *phosphor* on the inside front face of the vacuum display tube. The input signal deflects the trace vertically by the use of electrostatic deflection plates; the timebase generator of the scope scans the electron beam horizontally, again using electrostatic deflection. These scopes never achieved any great accuracy, typically having specs of 2% or 3% for both horizontal and vertical deflections.

Real-time scopes are not good for looking at waveforms with repetition rates of say 20 Hz or less; below this repetition rate the trace flickers and becomes dull. At these low repetition rates you have to resort to using a *scope hood*, a shaped opaque tube that fits onto the front of the scope at one end and around the viewer's eyes at the other. Thus low repetition-rate signals can be seen by lowering the effective ambient lighting level. Trace plots are made by putting a camera into its own scope hood in front of the screen.

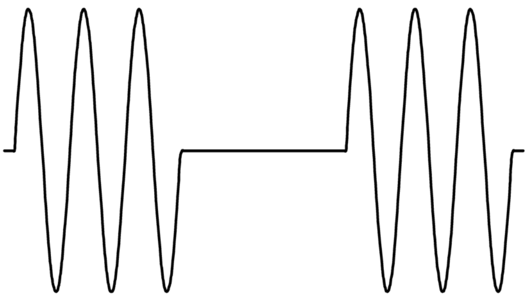
For real-time scopes it is essential to get a good trigger. Without a good trigger the trace will not be bright, or stable, and you won't be able to see what is happening in your circuit. If you are looking at simple sine/square and triangle waves then triggering will not be an issue. In the real world, though, you will be looking at arbitrary waveshapes, possibly with bursts of signal. These can be difficult to trigger on with a real-time scope. If you have a good expensive scope, then it may have a *trigger-holdoff* facility to help you to get a stable trigger.

A real-time scope has a sweep generator waiting to drive the beam across the screen.

[†] Tektronix TDS7104.

Once triggered it runs across the screen at a fixed rate. During this time the sweep generator is locked-out from re-triggering. It remains locked-out until the beam has been deflected back to the starting side of the display {left side}. It then obediently waits for a new trigger before sweeping again. The trigger-holdoff just increases the delay before the horizontal sweep generator is re-armed.

FIGURE 15.3A:



A waveform like this is easy to trigger on when you have all three cycles of the burst on the screen. If you want to zoom in on the first cycle, then the second or third cycles within the same burst may also cause triggers, confusing the displayed result. One answer is to turn up the trigger-holdoff so the sweep is not able to restart on the current burst of cycles. Provided the trigger is not re-armed until after the moving part of the burst, you can zoom into the waveform with ease.

If you want to look at detail after the trigger point, say on the third cycle of the above waveform, then you still need the trigger-holdoff to get the repetitive triggering, but now you need a *delayed timebase* as well. The delayed timebase allows you to zoom in on detail that occurs after the trigger point. The traditional way to show this is to have a mode where a section of the trace is made brighter. This brighter area is the area that will be expanded when you switch to the delayed timebase. It is sometimes difficult to set up a delayed timebase and frankly the modern approach would be to just use a *Digital Storage Oscilloscope* (DSO) instead.

There is one feature of real-time scopes that you should be very aware of. There is sometimes a control on the front which selects between *chop* and *alternate*. Modern real-time scopes have only one electron beam but two or more channels. To obtain two traces from one beam, the scope can either do one sweep on one channel, then one sweep on the other channel (alternate mode); or it can switch between the two channels as it moves across the screen horizontally (chop mode). Alternate mode can produce very strange effects at slow timebase speeds and chop mode is exceptionally difficult at fast timebase speeds. If given the option, select chop at slow timebase speeds and alternate for high timebase speeds. Often scopes will switch modes automatically according to the timebase setting.

Trigger hold-off and a delayed timebase are features that you have to specifically pay extra money for when buying a real-time scope. What you can't see from the spec, however, is the brightness. Suppose you are looking at one cycle of a continuous sinewave at 1 μs/div. This is not a difficult task and any real-time scope will give a clean bright display. Now suppose you want to look at the fast rising edge of a 10 kHz waveform. The waveform has a repetition rate of 10 kHz, but you are using the 5 ns/div range to view the edge. The trace is on for 50 ns and waiting for 100 μs. The trace is therefore only scanning for 0.05% of the cycle. A good scope will have enough brightness to show the edge and a bad scope will not.

Storage scopes:

Storage scopes come in two types: analog storage (tube storage) and digital storage. The tube storage type still exists in old test equipment, but they are so horrible I hope you never have to use one! You are getting more 'insider information' here as I used to design DSOs for many years.

The Digital Storage Oscilloscope (DSO) is now the king of the scope world, despite the fact that it was only patented in 1968,[3] and only became commercially available in the early 1970's. The DSO handles ridiculously low sweep speeds (>1000 s/div) up to higher effective sweep speeds than are possible on a real-time scope (<100 ps/div). The DSO therefore covers more time decades than the real-time scope, and has many added features. *It is essential that you master the operation of this vital piece of test equipment.*

The DSO can do automated measurements of risetime, overshoot, amplitude, frequency, period, duty-cycle, mean, RMS, cycle-RMS &c. It can be remotely controlled and used as a high-speed method of getting analog information into a computer. It can capture narrow pulses (glitches) that wouldn't even register on a real-time scope (because the glitch would not be bright compared to the rest of the trace).

The technology and functions available on the modern DSO are changing rapidly in the direction of ever increasing complexity. For this reason it is essential to read the manual to find out what whizzy features your particular equipment has. However, the basic features are common between types.

You should expect to be able to trigger on a waveform and to be able to move this trigger point from the left side of the screen to the middle and even over to the right of the screen. You are then seeing data which occurred *before* the trigger event occurred. Such exotic names as *pre-history* and *negative delay* are sometimes used, but the preferred term is *pre-trigger*.

There is nothing magical or clever about pre-trigger. The scope just continuously writes the acquired ADC data to RAM. When the trigger event occurs the hardware decides when to stop the acquisition. If the acquisition stops immediately then there is 100% pre-trigger data in the RAM. If the trigger edge is in the middle of the screen then that is 50% pre-trigger.

Actually, by the use of a delay line some expensive real-time scopes did enable you to see the front edge of a fast waveform. You could get perhaps 10% pre-trigger on the fastest timebase ranges, enabling all of the rising or falling edge of a waveform to be seen. At anything other than the top timebase ranges, however, they could only achieve trigger delay. On a DSO you should expect to be able to get anything from 0% to 100% pre-trigger on any timebase.

Unlike the real-time scope, the DSO does not need a stable trigger. The acquisition can be stopped at the end of a single sweep, displaying a stable picture of what was happening at the time. For more detailed examinations a stable trigger can be useful, but it is not nearly as important as for a real-time scope.

The DSO is the workhorse scope of the modern designer. No design project is complete without a print-out of the power supply rails as they start-up (power on) and shut down (power off). This is something that was almost impossible to do with a real-time scope and very messy with an analog storage scope {the old style Polaroid instant camera film used c.1985 was messy}.

[3] J.V. Werme, 'Chronological Trend Recorder with Updated Memory and CRT Display', *US Patent 3,406,387* (Oct 1968).

The key spec points for a DSO are:

- Maximum (single-shot) sampling rate.
- Maximum store (trace) length (in samples)
- Analog bandwidth
- Vertical resolution (bits)
- Vertical accuracy
- Glitch capture (minimum) width
- Automated measurement capability
- Analysis, such as FFT, and averaging
- Remote interface capability
- Data transfer and storage (floppy disk, hard disk, USB memory stick, Ethernet)
- Ease of use of user interface
- Display arrangements, trace colours, display in *roll mode*

The maximum sampling rate obviously limits how many points you can get on a fast edge; more points means more resolution. High sampling rates cost a lot of money so manufacturers “cheat” to push up the sampling rate. You will find a “single-shot” sample rate. This is the speed of the ADC. You will then find a *repetitive sampling rate*. On a repetitive signal it is possible to acquire the same signal many times before displaying it on the screen. The incoming signal will not be synchronised with the scope’s internal timebase, so each acquisition will start at a slightly different position on the waveform. By correctly interleaving the acquired data, a much higher effective sampling rate can be achieved. This is known as *equivalent time sampling* (ETS). It is not a nice mode to use because the trace can update slowly, but it does mean that you can get a repetitive sampling rate 10× to 100× faster than the single-shot sampling rate.

Maximum store length may not seem very important, but long stores do have important uses. The most obvious use of a long store is when you have an infrequent event and you are not sure how long the event will last. You set the store length to maximum so you can capture the whole region around the event and then you can zoom in on the part you want after you have captured the event. It is important to realise that the maximum sample rate of the scope only occurs on the top timebase. As soon as you switch down to the next lower timebase the sample rate has halved!

Glitch capture will show you any fast events that have occurred, but that doesn’t give additional horizontal resolution. A longer store gives you greater horizontal (time) resolution. Suppose you have a 1K store and you measure the frequency of one cycle of a waveform displayed on the screen. The timebase accuracy of the scope may be 10 ppm, but you have limited the resolution to ±2 parts in 1000 (±0.2%) by the use of the short store. Now it is true that the scope measurement algorithms may well interpolate the zero crossings for you to minimise the ±1 dot uncertainty at each end of the waveform, but even so, the reading is less accurate than it could be because of your mistake. The only reason you should ever use a short store length on a slow timebase setting is if the scope becomes too slow on the longer store setting.

You do have to be a bit careful about analog bandwidth. You might reasonably expect that the analog bandwidth was a fixed number for the scope, not least of which is because it is often written on the front panel in big print. Unfortunately some manufacturers reduce the analog bandwidth at slower timebases. They say that it helps to minimise ***aliasing***, whereas in fact they do it to make the scope look less noisy!

Another manufacturer’s ‘trick’ that you have to be aware of relates to the interpolation algorithm. When you get very few points on the displayed waveform due to the sample rate and the amount of horizontal zoom {expansion} used, the trace can get very "angular" with a linear dot join. In many respects this is good because you can see the real data points and it is obvious that the sample rate is too low.

One particular interpolation algorithm is extremely bad: the *sinc interpolation*. The problem is that the trace is completely smooth and continuous. There is no evidence that anything is wrong and yet you can be looking at a waveform where most of the data has been "invented". The sinc interpolator also creates nasty overshoots that are not present in the acquired data. My recommendation is that you always have interpolation set to linear to avoid problems with unreal data points.

Aliasing is not something that occurs on real-time scopes. Old time engineers used to worry about this a lot, but for modern engineers, brought up on DSOs, it is not a problem. A simple clue to an alias is that the trigger light is on, but the waveform is not stable on the screen. But do check that the scope is being triggered from the channel you are viewing! If you think the waveform you are looking at is an alias, it is a simple matter to select *glitch detect* {max-min mode; envelope mode}. If the waveform becomes a solid band then you know the scope was aliasing.

The DSO is the key tool used for noise debug work, and the key feature you need on the scope is averaging. If the DSO can't do averaging then it is virtually useless for noise debug work. Another key feature is the ability to do FFT analysis of the acquired data. Again this is essential for noise debug work because you can spot low level noise sources at definite frequencies; these may not be visible on the time domain waveform.

Sampling Scopes:

Don't confuse sampling scopes with DSOs (Digital Storage Oscilloscopes); they are entirely different things. Unfortunately some authors have used the incorrect expansion "digital sampling oscilloscope" for DSO, and this only helps to confuse the novice.

Sampling scopes are available in both digital and analog types, although the analog types have been out of production for many years. Regardless of whether they are of the analog or digital type, sampling scopes *only* work on repetitive signals. Their function is to view repetitive waveforms with frequencies into the tens of gigahertz or with risetimes of picoseconds.

A sampling scope works by taking one sample per trigger event, with each successive sampling point being delayed slightly more relative to the trigger point. This enables the signal waveform to be reconstructed by joining the sampled points together. The input of a sampling scope is a *sampling (diode) bridge*, which is pulsed by a very narrow pulse («1 ns). This pulse briefly connects the input signal to the sampling capacitor. The samples can only be taken at a maximum rate of about 100 kS/s because of the methods used to generate the fast pulse.

The key problem with this type of scope is the requirement for a trigger well before the edge being viewed. On older units it was required to have a trigger waveform up to 75 ns before the edge being viewed on the screen. If the incoming signal has a repetition rate of greater than about 7 MHz this is not such a problem. You can always look at the next waveform edge, or trigger on the opposite slope of the incoming waveform, providing that the waveform frequency and duty cycle are stable. However, if the waveform has a low repetition rate, say below 100 kHz, you need to somehow split the input signal, then put the main signal through a 75 ns delay line.

Splitting the signal is not good, since even the best 75 ns delay lines disperse a fast edge {slow it down and spread it out}. As an example, the [long since obsolete] Tektronix 113 50 Ω delay cable gives a 60 ns delay with 100 ps risetime and is the size of a large suitcase. Even a 100 ps risetime on the delay line is bad when you are looking at scopes with bandwidths up to 75 GHz (5 ps risetime). The requirement for an advanced trigger-pulse on sampling scopes makes them considerably more difficult to use than DSOs.

More modern sampling scopes [4] reduced the time skew requirement between the trigger pulse and the input signal down to 16 ns. Even so, the effect of the cable delay line still gave a noticeable worsening of the scope's performance.[5] Clock recovery circuits are then needed to get stable triggers.

Since modern DSOs are improving in bandwidth and sample rate, the need for sampling scopes is relegated to use beyond 15 GHz bandwidth (28 ps risetime).

15.4 Probes & Probing

Probing with a DMM is relatively straight forward. You just use two probes and measure wherever you want. The main difficulty comes when the circuit being probed gets upset {disrupted} by the capacitive load of the DMM. In this case, adding a resistor of between 100 Ω and 10 kΩ at the probe tip should stop the problem.

An alternative solution is to solder a small 0.25 W wire-ended resistor to the point you wish to probe. Don't leave the resistor leads more than 1 cm long though. For production testing requirements you should include a small resistor on the PCB in series with the test point, enabling technicians to more easily make the measurement.

Even "high impedance" scope probes (10 MΩ//15 pF) can cause problems to the circuit under test, and this trick of soldering a resistor to the point being probed is equally effective for solving that problem. Think of it this way: probing the circuit is *likely* to cause problems. If you are lucky, you may be able to get away without having to solder a resistor to the point being tested.

One trick is to make the test point a 1 mm diameter plated-thru hole. The probe tip is then centred nicely in the hole. A scope probe can also be placed in the hole, hands-free, the probe body pulling the tip into contact with the plated barrel in the hole. (Remove the probe's grabber hook first!)

Scopes typically have BNC connectors on the front, and often the inputs are labelled with something like 1 MΩ//10 pF. The scope input will look something like a 1 MΩ resistor in parallel with a 10 pF capacitor, perhaps up to 1 MHz. After that, the impedance becomes too complicated for this simple model.

It is not a trivial matter to connect a scope to the circuit under test. If you just put a BNC to 4 mm banana plug converter (or binding post) on the front of the scope you can hook up some ordinary test wire and measure something. This may just about be acceptable for low speed signals (<1 MHz) of high amplitude (>500 mV ptp), but what you will find is that for smaller and/or faster signals there will be a lot of noise picked up in the test leads. This banana plug connection scheme would be completely unusable to measure any sort of switched-mode supply for example.

What you need is *screened leads* and the usual answer is to use coaxial cable (coax).

[4] Tektronix TDS820. This was withdrawn from production at the end of 1999.

[5] The bandwidth of the (discontinued) TDS820 drops from 8GHz to 6GHz when the internal delay line is used

Unfortunately you will find that ordinary coax has a capacitance of perhaps 100 pF/m, so your 1 metre test lead plus scope input capacitance gives a loading on the circuit of at least 110 pF. This can easily be enough to make a circuit go unstable {oscillate} or at least make it ***ring***.

This brings us to 1:1 scope probes. These have a special low-capacitance lead of perhaps 30 pF/m [with resistive loss to reduce ringing; ≈100 Ω/m] so the load on the circuit may be around 40 pF. Nevertheless, the bandwidth achievable with this type of probe is strictly limited to the 1 MHz to 10 MHz area.

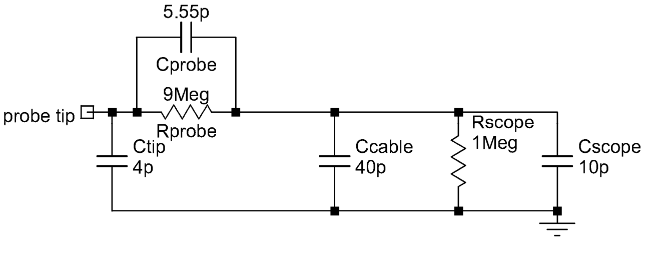
In reality 1:1 probes are not much use for measuring signals above perhaps 100 kHz. The standard probe to use is a 10:1 passive probe. A word of caution on terminology is appropriate at this point. A 10:1 probe makes the signal at the scope input 10× *smaller*. The probe is called a 10× probe or a ×10 (times ten) probe, however, because the signal measured on the screen has to be multiplied by 10 to correctly scale the result for what is occurring at the probe tip.

Some scope/probe combinations auto-detect the probe scaling factor and display the correct value. Some scopes have manually set probe scaling factors. You absolutely have to find out how the probe scaling works on your scope or your measurements could be wrong by a factor of 10×; career limiting performance!

To check for an auto-scaling probe, just unplug the probe and see if the volt/div reading on the display changes. To check if there are internally set scaling factors plug the probe onto the scope's 'cal pins' and see if the 1 V signal reads 1 V. Always do this with an unfamiliar scope in the same way that you would check the mirrors when you get into an unfamiliar car.

In its simplest form, a 10:1 probe consists of a 9 MΩ resistor with a variable shunt capacitor. Depending on the probe type, the capacitive input loading is reduced to something between 9 pF and 25 pF. Signal frequencies up to hundreds of megahertz can now be measured. The coax lead on the probe is still resistive [≈100 Ω/m] in order to improve the pulse response.

FIGURE 15.4A:



Expect to have to trim the scope probe for each individual channel on the scope. If you swap the probe to another channel, you may need to re-trim the probe. C_{probe} may be used to trim the frequency response of the probe. This is done by applying a 1 kHz square wave at the probe tip, using the scope cal pins, and adjusting for a flat topped response. Safety legislation (c. 1998) makes it difficult to have trimmers on the 'hot' end of the probe {the end connected to the high voltages} so the trimming is now more usually done at the scope end of the probe in a 'tail box' {the part that plugs onto the scope BNC}.

The scope probe is primarily a 9 MΩ resistor. If you connect the probe up to a high voltage, such as 230 V mains, before connecting the probe to the scope, you run the risk of getting a shock from the end of the tail box. This is why you must connect the probe to the scope before connecting the probe to the circuit under test.

In my equivalent circuit there is an ***earth*** (ground) symbol. This actually means the earth of the mains power system; the *circuit protective conductor* to use the formal terminology. On the majority of scopes, the case and the BNC's are connected solidly to mains earth. The BNCs should be able to sink at least 10 A earth fault current, but manufacturers interpret the standards differently and I would not like to guarantee that all manufacturers follow this rule. (All the Gould scopes I designed would take 25 A.)

Up until the 1980s it was common practice to remove the earth leads from scopes in order to ***float*** the scope up to some mains related potential. This was never a safe thing to do; you had to make sure not to touch the case of the scope whilst the circuit was energised. Some engineers, and some established companies, may never have grown out of this habit. With modern safety legislation, there is no place for this 'technique'.

In the first place, just removing the earth from a scope leaves the EMI filters nothing to connect to. There will be excess RF radiation emitted from the instrument. Even if you don't connect the scope to anything, there may be up to half the mains potential appearing on the case. This is due to the Class Y filter capacitors in the line filters. You will get an electric shock by just touching the case. Also, any circuit you connect the scope to will have this voltage injected into it.

Do not remove the earth/ground connection to a scope.
It is both illegal and unsafe.

If there were to be a serious industrial accident as a result of removing the earth from a scope (or indeed any other piece of ***Class I*** equipment) the manager who permitted it would undoubtedly be held personally liable. If somebody died, for example, a manager would probably end up in prison.

DEFINITION: Electrocution means "death by electricity".
Do not confuse 'electrocution' with 'electric shock'.

If you have a difficult measurement problem then use a proper measurement technique. Use a *differential probe*, use an isolated probe, use an isolated input scope, use an isolating transformer to float the circuit under test, but **don't** float the scope.

A 10:1 probe can have an input capacitance of 16 pF. Whilst 500 MHz bandwidth is achievable when driving the probe from a low impedance source (<20 Ω), the whole point of using the probe was that it was supposed to present a high impedance to the circuit being measured. 16 pF is not a high impedance at 100 MHz, it is ≈100 Ω.

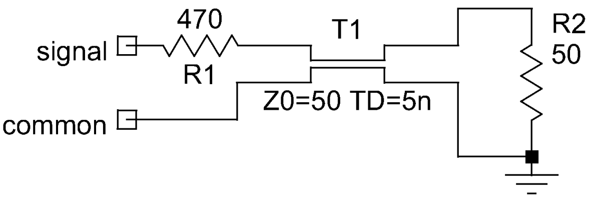
EX 15.4.1: Estimate the signal loss when measuring a sinewave at 100 MHz. The source 'looks like' 1 kΩ resistive at this frequency and the probe 'looks like' 10 MΩ//16 pF. Assume that the probe and measuring device have a totally flat frequency response up to at least 100 MHz.

It is possible to get probes with lower input capacitances; perhaps 10 pF at the probe tip for a 10:1 passive probe, maybe 4 pF for a 100:1 passive probe. The trouble is that there is very little signal left to look at. There are two solutions; a 'low impedance' passive probe and an active (FET) probe.

Active FET probes are great because you can get them in 1:1 versions with input resistances ≥100 kΩ and input capacitances ≤3 pF. Their main problem is price. They start at around $150, but can cost upwards of $1500, and they are not very robust, either mechanically or electrically. Breaking $1500 probes is a good way of losing friends and (adversely) influencing your career.

'Low impedance' passive probes have the virtue of giving very low circuit loading at say >30 MHz and you can make them yourself.

FIGURE 15.4B:



R1 feeds into a 50 Ω coax, terminated in the 50 Ω of the scope (R2). I use 1 nF in series with R1 so the probe is AC coupled. Both the resistor and the capacitor are 0603 surface mount types, supported on a small piece of FR4 printed circuit board. This makes a probe with very little ***aberration*** (<3% ptp) on a 200 ps rising edge, and bandwidth in excess of 2 GHz. This probe has a circuit loading of roughly 520 Ω, but this value is constant with frequency. The resulting measurements of high speed circuits are therefore lower in amplitude, but are the correct shape in the time domain.

This brings up another important point about probing in general. Ideally you would monitor the output of whatever it is you are working on, apply a probe somewhere else and see what happens. This tells you how much the probe is actually loading the circuit. Probes can make circuits do all sorts of 'interesting things'. They can both create or stop oscillations, they can inject extra noise, and they can cause offsets. You have to see what they are doing by having additional monitors in place. It may be that you therefore need at least two probes. One of the probes may be placed further down the amplifier chain, allowing you to monitor the effect of a probe placed earlier in the amplifier chain.

Another trick, which is only useful for looking at large signals, is to put a 1 pF capacitor in series with a 10:1 probe, right at the tip. Use a wire-ended capacitor and wrap one of its wires around the probe tip (probe grabber clip removed). The other end is used to probe the circuit. This home-made probe now has a capacitance of around 1 pF although its calibration is uncertain and it is AC coupled at quite a high frequency.

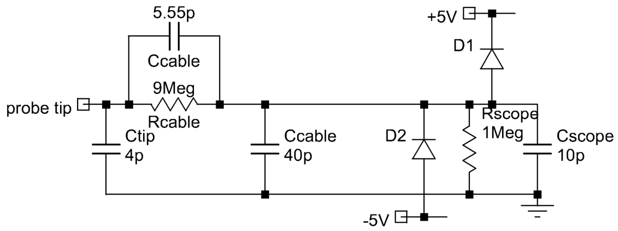
EX 15.4.2: A basic 10:1 probe has an input impedance of 10 MΩ//12 pF.

a) What is the AC coupling corner frequency using 1 pF into the probe tip?
b) Assuming a 2 mV/div sensitivity at the scope input, what is the sensitivity at the probe tip when the 1 pF capacitor is used?

All scopes can be non-destructively overloaded, but the speed of recovery varies dramatically with manufacturer, model type and volt/div range. If the waveform goes off the top and/or bottom of the screen, the result is no longer defined [unless explicitly stated in the manufacturer's data sheet].

It is normally fairly safe to have one screen of overload at low frequencies, say <1 MHz on a 100 MHz scope. However, if you try this at the highest speed the scope is capable of, you will certainly see an error. You can investigate this effect by taking a fast edge which is perhaps 6 or 8 div in amplitude and shifting it almost completely off the screen. Does the corner shape change with screen position? [Probably.] If you want to make valid measurements you should not be driving the scope outside its linear range, or else you will have to separately characterise its performance when you do so.

FIGURE 15.4C:



This is the same 10:1 probe circuit as before, but now some clamp diodes have been shown within the scope input. This is an over-simplified circuit, but a similar effect will happen in a real scope.

***EX 15.4.3:** You are measuring on the 50 mV/div range, with the trace shifted down to the bottom of the screen. The effective sensitivity of the scope is 500 mV/div because of the ×10 probe being used. The scope is not overloaded with the 8 div (4 V) 10 kHz signal being observed. Now the probe is moved to a new point in the circuit which has the same wave-shape, except that it has large narrow spikes on the edges which extend up to 100 V. The scope is rated to 400 V peak, so you are not bothered by these spikes, particularly since the 10:1 probe limits the signal seen by the scope input. Neglecting problems with the overload recovery of the amplifier (not shown in the sub-circuit), what is the problem with the probe/scope combination?

The clamp voltages in that example were generous at ±5 V. The input may be clamped at ±1 V or less. High bandwidth scopes (>100 MHz) are more difficult to make than lower bandwidth scopes, since high speed semiconductors are more electrically ‘fragile’ than their lower speed counterparts. Because of this, it is quite possible that a lower speed scope will recover from certain overloads faster than a higher speed scope. One factor is the voltage to which the clamp diodes are held. Because the high speed devices are more fragile, the input clamp diodes on a high bandwidth scope are likely to be held to a lower voltage than that used on a lower bandwidth scope. Again the only ways to know for sure are either to test it or to ask the manufacturer.

There are specific tests that you can and should do before believing the result of a probing operation. The first test is to connect the probe tip to the probe ground lead at the point to which the ground lead is attached to the circuit. This should give a low level of noise. If it does not then there are two possible reasons that need to be separated out. Firstly, the noise could be due to radiated pickup. To test for this, connect the probe ground lead to the probe tip, but with neither actually touching the circuit under test. If there is still noise, it is either radiated or magnetic pickup.

One solution to this problem is to shorten the ground lead and/or the probe tip. If there is a removable grabber clip on the probe then remove it. Now you will see a small pointy tip and a grounded ring. Put a coil of tinned copper wire around the grounded ring with a short extended tail. You now can probe with much shorter ground and probe leads, giving proportionately less radiated pickup.

If you don’t get any noise on the radiated pickup test then you have *common-mode current noise*. One way to reduce this noise is to provide another path for the common-mode current to go down, rather than through the probe ground lead. Connect earthing wires from the equipment under test to the scope, perhaps using adjacent unused channels on the scope as grounding points.

Another effective technique is to reduce the length of the probe ground lead as suggested above. The reason this works is actually rather complicated. Obviously a 1 m scope probe has a considerable inductance. This inductance is not at all relevant to the measurement though. The inductance of the 5 cm of ground wire near the probe body is critical however. Once the signal gets into the coaxial part of the cable, the cable acts as a 1:1 transformer and couples the voltage drop across the outer sheath onto the inner conductor. The effect of this tight coupling is to heavily attenuate the common-mode noise.

All coaxial cables achieve this tight coupling above a few kilohertz. All coaxial cables have negligible coupling when the frequency of the screen current drops below a hundred hertz or so, the exact point being called the *shield cutoff frequency*. Below ten hertz, the voltage drop in the probe ground lead is not coupled to the signal path in any way. Thus voltage drop in the probe ground lead appears as an unwanted noise signal at the mains frequency and below.

Only when you have done these tests can you be assured that you have made a correct measurement. If you just probe the signal directly without these tests then a “correct” signal could actually be incorrect because the noise voltage happened to add in a helpful manner. This statement is particularly true in the immediate vicinity of switched-mode power supply components; the field pickup can easily be 10× greater than any signal you are trying to measure.

The two biggest sources of gross errors when using scopes are incorrect probe gain setting and incorrect probe compensation. Some scopes have automatic probe sensing when used with special probes. The volts/div setting of the scope changes for you when the probe is connected. Some people miss this and apply another scaling factor for themselves. Other users think the scope is sorting it all out, but use the wrong type of probe without the sense circuitry in, so the scope doesn’t realise the probe is there. Some scopes have scaling factors that are just manually entered and users switch from 1:1 cable connections to 10:1 probes, forgetting to change the scaling factors. Maybe somebody else has been using the scope and left these scaling factors set.

Just be alert and know your test equipment. Then touch the probe to the calibrator pins on the scope’s front panel as a final check to make sure that you don’t get readings that are out by a factor of ten! Oh, and when you touch the probe to the cal pins, make sure the pulse response is flat. It may not be the scope that causes the pulse response to be unflat, it may be the adjustment of the probe. Always assume that the pulse response of a probe needs to be re-adjusted for the particular scope channel you are using!

Fail to take these basic precautions and you will both look and feel like the idiot you have been. Presenting readings that are out by a factor of ×10, or with pulse response errors due to incorrect probe adjustment, is not acceptable for a graduate of this course.

15.5 The Spectrum Analyser

Get the manual and skim over it to find and read the good parts. I cannot overstress this point! Did you ever hear of a pilot getting into an unfamiliar plane and just flying it?

You may not be working on a >50 MHz circuit, so why should you be using a >2 GHz spectrum analyser? There is one very specific reason for this. Spectrum analysers are very sensitive detectors of system noise. You can hook a spectrum analyser onto an amplifier output and see exactly what is going on. General digital ‘bus noise’ will show up on a spectrum analyser, but it won’t show up nicely on a scope because there may be no definite repetition rate.

Having said that, if you get a modern digital storage scope with built-in FFT analysis, you effectively get a spectrum analyser which covers the band from DC up to half the sampling rate of the scope. Ordinary spectrum analysers designed for multi-gigahertz performance sometimes do not function at all below a few tens of megahertz. This is especially true on older models, or modern versions where LF operation is an optional extra which somebody didn't care to pay for.

LF circuits can oscillate at hundreds of megahertz without you knowing. Generally speaking, opamps oscillate at modest frequencies (say a few kilohertz to a few tens of megahertz) when you do incorrect things to them. These oscillations will usually show up on a good scope. But any circuits with transistors in can oscillate at unreasonably high frequencies. You must be capable of looking for spurious oscillations up to half the f_t of the transistors you are using. Discrete silicon transistors are readily available with f_t values of up to 32 GHz (Infineon BFP520). These values may increase as technology advances, but other types of device are also available. Silicon-germanium (SiGe) discrete transistors are available with f_t ratings to 65 GHz (Infineon BFP620). Given any slightest opportunity these will oscillate and you have to establish that they are not doing so, or cure the oscillation completely if they are.

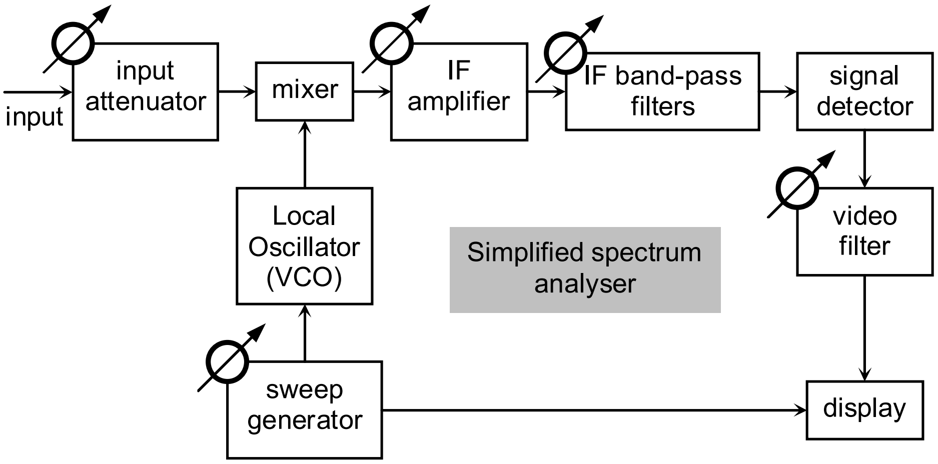
The interesting thing is that with a spectrum analyser you can also see where the circuit is getting close to instability; the noise floor will peak at this point. Excessive noise peaking (>20 dB) should be investigated and eliminated if possible (given time & cost constraints) by better decoupling or better layout.

Spectrum analysers have 50 Ω inputs, which can be quite inconvenient for noise debug work, especially as they are often type-N inputs. Also, they are usually very fragile electrically, many being easily damaged by static electricity and most being damaged by voltages greater than ±5 V. One solution to this problem is to put a type-N to BNC adapter on the front of the analyser and use an ordinary 10:1 scope probe for diagnostic work. If you are only looking for oscillations, this technique will give you a good way of probing around without damaging the spectrum analyser. The amplitude values will all be wrong, but then again, all you are looking for is the presence or absence of an oscillation.

If you directly connect your scope probe to an active component or a large signal you may inadvertently *cause* an oscillation. The probe clip is an antenna {aerial} for signals above say 30 MHz and also a capacitive load. What you can do instead is to clip the probe onto a nearby ground point. That was not a printing error. Put the signal pickup of the probe onto a ground point {signal 0 V; earth} in the amplifier that you are working on. The spectrum analyser is so sensitive that it will still register if there is an oscillation going on anywhere nearby. This gives the best chance of not loading the circuit or creating a false oscillation. When the circuit passes this test then you can go on to probing the actual signal using a passive low-impedance probe.

Not all spectrum analysers are equal. If you just look at the "banner spec" features of price and bandwidth then you may well consider that the cheaper one represents excellent value for money. Beware! Cheap spectrum analysers may have very poor dynamic performance. You need to check the small print in the specs. Compare the *displayed average noise level* (see ***Noise Figure***), the harmonic distortion, the *phase noise*, the *shape factor* of the IF filters, the number of IF filters, and the type of signal detector(s) as a minimal comparison.

FIGURE 15.5A:



A spectrum analyser is a complex piece of equipment and has several controls that have to be set correctly in order to get a valid measurement. The manual will help you to understand these controls. The major user controls have been shown on the simplified block diagram as circles with arrows through them. All of the settings are interlocked by default to stop you doing silly things, but these interlocks can be partially or fully by-passed by advanced users to get better performance.

The first thing you have to do is to set the analyser up for the amplitude of the input signal you are applying. The spectrum analyser has a *reference level* which is the top horizontal graticule line. Ordinarily you would set the reference level just above the peak of the applied signal in order to get optimum measurement resolution.

The reference level is internally set using two controls: the *input attenuator* and the *Intermediate Frequency* (IF) amplifier gain control. There are multiple setting of the input attenuator and IF gain that give the same reference level on the display. For any given reference level, less input attenuation gives better noise performance, but worse ***intermodulation*** and harmonic distortion performance. The spectrum analyser will automatically use a compromise setting to balance these two effects.

The IF filters select different *resolution bandwidths*. If you need to look at the detail in a swept response, reduce the IF bandwidth. As the IF bandwidth is reduced either or both of the sweep speed and the sweep width [MHz/div] need to be reduced. The system will decide for you and you may need to adjust the setting manually.

These IF filters are not perfect ***brickwall*** filters. They have performance specified in terms of ***shape factor***; a measure of the steepness of the attenuation characteristic outside the pass-band. This spec may also be referred to as *selectivity*; the better the selectivity, the lower the number.

$$\boxed{\text{Shape Factor} = \frac{\text{-60 dB bandwidth}}{\text{- 3 dB bandwidth}}}$$

If the IF filter is set to 1 kHz [3 dB] bandwidth and it has a shape factor of 11 [a typical value for a good spectrum analyser] then the attenuation will not reach 60 dB until the frequency is $1\,\text{kHz}\cdot\frac{11}{2} = 5.5\,\text{kHz}$ away from the centre frequency of the filter. You see this effect when looking at pure sinusoidal oscillators. Rather than seeing a single spectral line, as you reduce the frequency span you see a spread-out bell-shaped response. This could be due to the ***phase noise*** on the oscillator, but it could also be the IF filter response [with some contribution from the phase noise of the spectrum analyser's local oscillator]. In order to make a reasonable measurement you have to 'zoom in' by reducing both the sweep span and the resolution bandwidth. The clue is how smooth the curve is; a noise-free smooth curve will be due to the IF filters whereas a ragged response will be due to phase noise in the source and/or the spectrum analyser. Phase noise looks noisy!

Modern spectrum analysers use digital filters for the lower bandwidth IF filters, and these easily achieve a shape factor of 5; a much better selectivity than the 11 typically achieved using analog filters.

When comparing the displayed average noise level (DANL) spec, be very careful that the comparison is done at the same resolution bandwidth, or that you correct for the resolution bandwidth by 10 dB for every decade of resolution bandwidth above 1 Hz. Typical good figures for displayed average noise level are –150 dBm in 1 Hz resolution bandwidth, –140 dBm in 10 Hz RBW, –130 dBm in 100 Hz RBW, etc. Beware of numbers like –167 dBm in 1 Hz bandwidth as these are only achievable when a preamp is used in front of the mixer. Note that a DANL of –150 dBm/Hz means a noise figure of 24 dB, which is why preamps can give such good noise performance improvement.

Just what the "signal detector" does is an interesting question. You would hope that this detector would give an RMS reading of the signal within the IF band. However, if you deliberately put two equal magnitude sinusoidal signals closer than the IF bandwidth, the result is a fuzzy band, perhaps as much as 6 dB higher and lower than the original signals. When filtered, the original amplitude is seen. Thus a simple detector on a spectrum analyser does not measure RMS values at all. It is RMS calibrated for sinusoids, but it reads low on Gaussian noise by about 2.5 dB.

The video filter is another tricky control. It averages out the noise on the display so that particular spectral lines can be seen more readily. Too much filtering and the specific lines of interest will be averaged out as well. Again this filtering interacts with the sweep speed and the sweep range settings.

A safer alternative to the video filter is digital averaging from one sweep to the next. This does not affect the other settings, but does reduce the noise floor, allowing constant small signals to be seen. The only warning here is that the signal does have to be constant. If the signal is moving about in frequency then it will also be reduced in amplitude by the averaging process.

When viewing a spectrum analyser display, there may be significant harmonics and spurious signals displayed. A good test to see if this distortion is due to the *input mixer* is to put an external passive attenuator on the front of the spectrum analyser. A 6 dB ***pad*** is more than sufficient. All true signals should drop by 6 dB. If some drop by significantly more than 6 dB, ideally by 12 dB or 18 dB, then you know that the spectrum analyser's input mixer was being overloaded and producing significant harmonic distortion or intermodulation distortion products. This test is a very powerful check on the signal level applied to the spectrum analyser's input mixer. To get the lowest possible internal distortion in the spectrum analyser, always attenuate the signal until the harmonics or intermodulation products are sitting near to the noise floor of the instrument.

When you get more skilled you will be able to do the same trick using the input attenuator on the spectrum analyser itself. I say when you get more skilled because you need to change the input attenuator and not the IF amplifier gain. A more complicated spectrum analyser may obscure exactly which settings are actually being changed. Note that high frequency (>100 MHz) harmonic distortion and intermodulation distortion on a spectrum analyser are *only* due to the input mixer; the IF stages run at fixed frequencies and cannot produce harmonic distortion.

Lower frequency harmonic distortion is a different matter though. It is now common for spectrum analysers to be usable all the way down to a few tens of kilohertz. This can be achieved by mixing the signal up to some gigahertz intermediate frequency as the first step in the acquisition process. In this case it is easy for the signal harmonics to fit within the bandwidth of the first IF stage; harmonic distortion in the first IF stage can therefore contribute to the harmonic and intermodulation distortion. It is for this reason that spectrum analysers often give poor or unspecified distortion performance below 1 MHz.

Traceability to national standards of harmonic distortion at frequencies beyond 100 kHz is not possible. In any case, RF signal generators generally produce harmonic distortions worse than –60 dBc (Agilent N9310A is –30 dBc). The only way to obtain a pure sinusoidal signal at megahertz frequencies is therefore to filter the output of a standard generator. The problem then arises as to how to prove that the filter itself is not generating significant amounts of harmonic distortion.

The aim of the exercise is to make measurements at extremely low levels of harmonic distortion; levels which are beyond the capability of generators, and beyond the range of spectrum analysers as well. The answer lies in the use of an absolute harmonic filter.[6] A simple length of open-circuit coaxial cable can be used as an absolute harmonic filter at one frequency. The idea is that it is used as a ***quarter-wave transformer*** at the frequency of one of the harmonics. The quarter-wave transformer converts the open-circuit load into a short-circuit at its input, thereby attenuating the incoming signal.

For a 1 MHz fundamental, the second harmonic is at 2 MHz and the coaxial cable needs to have a propagation delay of $\frac{1}{4\times 2\,\mathrm{MHz}} = 125\,\mathrm{ns}$. At roughly 5 ns/m this amounts to 25 m of cable. The actual attenuation achieved depends on the attenuation loss in the cable. As a practical example, RG58 coax used for a 1.1 MHz fundamental produced an attenuation of 20 dB at the second harmonic.

Now an absolute harmonic filter like this one is guaranteed to not introduce its own harmonic distortions, so this type of filter can be used to resolve uncertainties about whether or not a particular component is producing harmonic distortions. One proceeds as follows:

Buy or make a fixed frequency low-pass filter which has adequate attenuation to remove the harmonics from the signal generator you wish to use. Suppose your generator produces –40 dBc harmonics and you require better than –120 dBc output harmonics; you therefore require a filter which attenuates at the second harmonic by better than 80 dB. This level of attenuation will require at least an 8-pole filter. You now think you

[6] L.O. Green, 'Absolute Harmonic Filter for RF', in *Electronics World* (Highbury Business Communications), Mar 2004, p. 26.

should have an output harmonic level which is beyond the measurement capability of your spectrum analyser.

The next step is to buy or make a high-pass filter to remove most of the fundamental from the signal in order to improve the harmonic measurement capability of your spectrum analyser. The spectrum analyser should measure to better than –70 dBc, so this high-pass filter only needs to attenuate the fundamental by say 50 dB. You now have a system consisting of a signal generator, a low-pass filter, a high-pass filter and a spectrum analyser. The trouble is that you cannot calibrate the system because all four components are producing unknown amounts of harmonic distortion.

Take the absolute harmonic filter and shunt it across each junction in turn down the chain. At the signal generator output it produces no change on the spectrum analyser display. Good: the signal generator harmonic is not causing a problem. At the output of the low-pass filter it produces no change. Good: The low-pass filter is sufficiently linear. At the output of the high-pass filter it produces a reduction. Bad: We couldn't see any reduction of harmonics out of this filter by reducing the harmonics going in, but there are still harmonics visible; these must be due to internally generated distortion. Using the absolute filter you can qualify each component in the chain up to the required level of uncertainty. You have then made a measurement system with harmonic distortion capability well beyond what is commercially available.

Another confidence building test you can use involves a simple inline attenuator. Suppose you do a test with a 6 dB pad first at the input and then at the output of one of the filters. The rest of the components in the chain will still see the same fundamental signal level. The filter, on the other hand sees a change of 6 dB in signal level. If it is contributing a significant amount of harmonic distortion then there will be a noticeable change between the two tests. This is a powerful check for each of the filters.

15.6 International Standards

As far as the world of physics is concerned there are 5 *base units* immediately related to electronics: kilogram (Mass), metre (Length), second (Time), ampere (Current) and degrees Kelvin (Temperature). Defined standards for four of these allow derivation of the main electrical units. This is the ***SI*** system, existing since 1960. The Ampère is a base unit, but we need definitions of the other electrical quantities: primarily the volt, the ohm, the watt and the hertz.

Hertz are cycles per second. Watts are joules per second. The joule comes from physics: $\text{work done} = \text{force} \times \text{distance}$.

Newton's Second Law gives $\text{force} = \text{mass} \times \text{acceleration}$. Acceleration is distance per time per time.

Dimensionally the watt is therefore: $[Watt] = \frac{[Joule]}{T} = \frac{[Newton] \cdot L}{T} = M \cdot L^2 \cdot T^{-3}$

Then $P = I^2 R$ gives the ohm and $P = V \cdot I$ gives the volt.

This theory relates electrical units back to the SI base units. This does not necessarily make a convenient *primary standard*, or an accurate and reproducible one either. The key thing about primary standards is the ability to have several of them at different worldwide locations without allowing significant drift between them.

The SI definition of the Ampère is that current which produces a force of 0.2 μN/m in two infinitely long, straight, parallel wires of negligible cross-section held 1 m apart in a vacuum! As you can imagine this is not easy to measure.

In practice the ohm and the watt were used as the primary units; the ohm being evaluated from measurements of a calculable capacitor. From the ohm and the watt, the volt was agreed in terms of banks of Weston standard cells. But technology has moved on and the Weston standard cell, patented in 1891, is no longer the prime source of voltage standards.

The *AC Josephson Effect* has been known about since 1962. It occurs at the very low temperatures required for superconductivity and involves a current tunnelling through an insulating layer. The net result is that if the device, a *Josephson junction*, is illuminated by millimetre wave energy, the voltage across the junction changes in integer multiples of the frequency divided by the Josephson constant $\left[K_J = {}^{2e}\!/_{h}\right]$. The amount is not very great though, so it is necessary to put several thousand of these junctions in series in order to get an output voltage above 1 V. NPL uses devices containing at least 3000 junctions, activated by an 87 GHz source.

The Josephson constant was itself calibrated in SI units by a variety of techniques to an accuracy of around 0.83 ppm. Whilst the uncertainty of the Josephson constant was 'known' to only 0.83 ppm, it was established that the inter-comparison measurements could be done with an unprecedented accuracy of a few parts in 10^9. For this reason, on January 1st 1990, the Josephson constant was *defined* to have an exact value. The ±0.8 ppm uncertainty was set aside, by agreement. This gives a very reproducible and repeatable standard which is arguably not 'accurate' relative to the fundamental SI units, but which is more suitable than the previous standard.[7]

In the same way, the ohm was redefined in terms of the *quantum hall effect* with an agreed value of the von Klitzing constant $\left[R_K = \frac{h}{e^2}\right]$. The common uncertainty of ±0.4 ppm was also set aside. This business of getting fundamental constants from atomic standards rather than *artefact standards*[†] is known as *quantum metrology*. At present the only SI base unit that remains as an artefact standard is the kilogram.

The result of these changes was that as from 1990 the volt and the ohm were changed at all national metrology institutes. At the National Physical Laboratory (NPL) in the UK, the volt was decreased by 8.06 ppm relative to its previous value. NPL traceable resistances were decreased by 1.61 ppm. In the US, NIST traceable measured voltages decreased by 9.26 ppm and NIST traceable measured resistances decreased by 1.69 ppm.

Now we have an 'inversion' in the standards. Whilst the SI unit is the ampere, with the volt as a derived unit, we now have the situation where the volt is the very accurately and reproducibly specified unit. It is not an 'exact value' as far as an SI unit is concerned, but it is considered exact for the purposes of comparison.

The agreement between voltages at the level of national metrology institutes is now better than 1 part in 10^9. But this performance is unattainable by ordinary calibration laboratories. Even with the latest generation of voltage standards based on ultra-precision zener references, 1 year stability of 1 ppm with predictability to ±0.1 ppm

[7] Booklet, *Direct Current and Low Frequency Electrical Measurements* (NPL - National Physical Laboratory, UK, 1996).

[†] An *artefact* is something made using human work or skill. For 70 years prior to 1960 the metre was defined by two lines marked on a platinum/iridium bar.

for voltage is only just possible.[8] It is also found that for noise below 0.01 Hz, standard cells can exhibit less than 10× the noise of zener references.

Time and frequency are the most accurately reproducible SI units. In the UK there is MSF, the 60 kHz transmitter at Anthorn whose carrier frequency is accurate to 2 parts in 10^{12}. The US has a similar service, WWVB, with a quoted accuracy of 10 parts in 10^{12}. The low frequency transmission does not suffer from fading and interference as much as higher frequency services. Therefore you should expect these 60 kHz transmitters to be available for many years to come. The NIST short-wave transmissions WWV and WWVH in the band between 2.5 MHz and 20 MHz are also transmitted with the 10 parts in 10^{12} spec, but the *received* accuracy is not expected to be better than 0.1 ppm due to propagation effects.

15.7 The Prime Company Standard

It is usual for a company to have a measuring device, or some artefact standard, sent out to a certified calibration laboratory for annual calibration. The company then references all measurements to this prime standard and can therefore state that it has *traceability* to national standards with a defined uncertainty.

It is true that local calibration laboratories typically collect equipment themselves, ensuring that the transportation is done in a controlled manner. However, it is also true that the greatest likelihood of failure for a piece of equipment occurs when it is either switched on/off or when it is transported. These situations encourage failure, and in any case failure is also possible at other times.

What does failure of the prime standard mean? In this circumstance it means that the standard is not functioning within its expected level of accuracy. The only way to ascertain this is to measure the prime standard. If you wish to have a satisfactory calibration scheme, *you must never have only one prime standard*. The very minimum is to have two standards, then at least you can know that there is a problem. Three or more would tell you which one was at fault.

An example will clarify the situation. Suppose your company wishes to have a standard for resistance at 10 kΩ. There are two ways of doing this. Either the company can have a 10 kΩ resistor which is sent out for calibration, or the company can have a resistance measuring device such as a DMM sent out for calibration.

Suppose it has been decided to use a DMM as the company standard. It is sent out for calibration and comes back with a calibration sticker and a certificate of calibration. Wonderful. The company now ships its product, a poly-morphic de-fractaliser,[†] each of which is adjusted and measured on the DMM before it is shipped to the customers. Because each unit is being adjusted for optimum calibration against the prime standard, there is no way to know if the prime standard itself is still in spec. If it drifted off, the first indication might be that the pots on the production units could not be adjusted to give the correct reading. This is an unsatisfactory calibration scheme.

This next example shows how a calibration system should be run. You have a DMM and a fixed resistor of lesser accuracy. You measure the resistor with the DMM and record the value. You do this weekly. Over time you establish the repeatability of the

[8] Fluke 7010N Nanoscan Volt maintenance systems

[†] Invented product type.

measurements between the two units. Now you have two independent units telling you the same answer. If either suddenly changes, you know that there is a problem and you can do something to remedy the situation.

Just before the DMM is sent to the calibration laboratory you measure the resistor, recording its value as usual. You send the DMM for calibration and you make sure that the test house gives you readings before and after any adjustments that they may make. This is very important. When the DMM comes back from the calibration laboratory you re-measure the resistor with the DMM and record the value. You should get a reading that is consistent with the one week drift of the resistor/DMM pair, provided you take into account any change in the DMM calibration performed at the calibration laboratory. If the readings do not tie up, then there is a problem to be dealt with. This scheme gives a workable and robust calibration system. If the prime standard is damaged in transit, or just fails randomly, the error will be rapidly spotted.

The key to a robust calibration system is having more than one measuring device or measured standard. You can have two of the same measuring instrument type and still have a useful calibration scheme. Still using the resistance example, the scheme would use a resistor to *transfer* the accuracy from one DMM to the other. You only have to have one DMM calibrated at an external calibration laboratory. Before it is sent out, you measure the resistor first with it, and then with the second DMM. The resistor could be considered not to drift at all during this transfer measurement, but TC and noise effects would definitely have to be considered. The uncertainty of the transfer would be due to the finite resolution of each DMM, any change of temperature and any noise on the readings.

Which should be used as a prime standard; a resistor or a DMM? Fundamentally a standard should be very stable, very robust and very reliable. A standard should ideally also be very simple. A simple thing is less likely to go wrong or to have any sort of weird parametric problem. A fixed resistor is an excellent standard. There is very little that can go wrong with it. It always has the possibility of being more accurate than an equivalent measuring device.

You should get a lower uncertainty with a resistor as a prime standard. On the other hand, the DMM can give measurements across a broad range of resistances. It can also be put into an automated system and fully calibrated without involving staff. So the fixed resistor gives a more accurate calibration path, but it may be more expensive to use (if you are covering several decade ranges of resistance).

The general rule is that the simplest device *can* be made the most accurate. Whenever there is change of amplitude, change of frequency, change of load, change of something, then there is an associated additional uncertainty. It therefore stands to reason that the simplest device can be made the most accurate. A fixed resistor has but one function. The DMM which is measuring it has to handle a variety of different resistances and therefore has to be a more general solution.

Clearly this does not mean that any particular resistor will be more accurate than any particular DMM, but it does mean that in a contest to make the most accurate device, the resistor would win against a DMM, and a device which could only measure one value of resistance would probably win against a variable resistor.

A measurement has to be made against a similar thing; a voltage against a voltage, a frequency against a frequency (or a time). A possible exception is a *self-calibrating standard*. Certain ratio standards can be calibrated against themselves, giving a calculable performance rather than requiring a measured performance. The two key examples of this are the ***Kelvin-Varley*** [†] divider and the ***Hamon*** [‡] transfer standard. These have a calibration procedure to follow which checks out the individual parts one by one. Only when each of these parts is verified is the device calibrated to a defined level of uncertainty.

So far it has been assumed that "everyone knows" what *calibration* means. Everyone does know, but their agreed definitions vary! Consider a laboratory standard 10 kΩ resistor. Calibration of this consists of measuring it and reporting its value, traceable to national standards, with some defined uncertainty. Adjustment is generally not possible.

Now consider a more complicated piece of equipment such as a DMM. Calibration of this could consist of either measuring it and reporting its value, or re-adjusting it to optimise its value before reporting its uncertainty. Due to different terminology in different parts of the world, and at different laboratories, it is essential to agree in writing what is required for your calibration. There are several different types of calibration, some of which will not be possible on simple items like standard resistors.

Calibration types:

1) Measuring and reporting the value(s) to be within some stated limit(s), such as the manufacturer's specification, under specific stated conditions.
2) Measuring and reporting the actual measured value(s), with a defined uncertainty, under specific stated conditions.
3) Optimising the settings and reporting the measured value(s) after adjustment, with a defined uncertainty, under specific stated conditions.
4) Measuring and reporting the measured value(s), with a defined uncertainty, both before and after optimising the settings. {Type 2 followed by type 3}.

It should be evident that these calibrations get progressively more expensive as you go down the list. It is usual to need to optimise settings after the first year of use of a complex piece of equipment. After that, subsequent adjustments may do more harm than good. Some types of surface mount capacitive trimmers, for example, can be worn out by even a few cycles of adjustment.

[†] eg Fluke 720A Kelvin-Varley divider

[‡] eg Guildline 9350 Hamon resistance transfer standard

CH16: measurement techniques

16.1 Measurement Uncertainties

The tolerancing that has been given in an earlier chapter is that which should be applied to a design. But in the world of *metrology* (the science of measurement) there exists a laid down procedure for accredited calibration laboratories which you are required to follow. The guidelines are stated in the ISO document *Guide to the Expression of Uncertainty in Measurement*, a very bulky document. Even NIST Technical Note 1297: Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, a summarised version of the ISO guide, extends to more than 20 pages.

The text that follows is not a summary of the summary for use by staff at calibration laboratories. It is an explanation of what the calibration laboratories are talking about when they tell you how accurate your equipment is. If you are actually going to be calculating these uncertainties for yourself, you will need to at least read the NIST version of the guidelines.

In normal use you would expect to measure a voltage with a DVM, look up the spec of the DVM and make some sort of definitive statement. You might say:

"The voltage measured was +10.0011 V DC. The DVM has a one year spec of ±30 ppm of reading ±2 digits over the temperature range 23°C ± 4°C. The temperature was 22°C (±1°C) and the DVM was calibrated 4 months ago, so the *actual* value of the voltage is +10.0011 V ± 0.0005 V."

Having been careful to zero the meter leads at the measurement point and reduce ***thermal EMF*** effects by sensible use of the connecting lead types, you would have been very confident in your measurement. If you then tried another DVM and got an answer that agreed within the combined uncertainties of the equipment you would have been dead certain. However, if you got a certified calibration laboratory to measure the value, they might have come back with a value of +10.00105 V with an estimated uncertainty of ±0.00019 V at a 95% confidence level. Why are they so un-confident in their results?

The problems come about because of using a Gaussian (Normal) distribution model for measurements. If, for example, you modelled the distribution of resistor values as Gaussian you would find that you could never be totally sure that the value was less than a certain limit. Take the case of a 10K resistor with a ±5% tolerance. If the distribution is considered to be Gaussian with ±5% representing the $\pm 3\sigma$ limits then in theory the value *can* be greater than ±5%.

***EX 16.1.1:** A mathematician represents a 10K ±5% resistor's distribution as Gaussian, with mean μ=10K and standard deviation σ =167 Ω. According to this model, what is the chance that the resistor is outside of the range:

a) ±6% ?
b) ±8% ?
c) ±10% ?

For the calculation of the overall uncertainty of a reported measurement there are two types of uncertainty labelled Type *A* and Type *B*. Type *A* uncertainties are evaluated by statistical means. For example, on measurements above say 30 MHz in 50 Ω systems there is a contribution due to the connectors. Every time you make the connection you get a slightly different answer. To evaluate this as an uncertainty you would do an experiment of making and breaking the connection, perhaps 5 times or more, and you would calculate the mean and variance of the measurements. This would give a Type *A* uncertainty. The term *random* uncertainty is sometimes used for a Type *A* uncertainty, but is non-preferred and strongly discouraged.

Type *B* uncertainty contributions are *unknown* systematic errors. This is an important point. If you knew that using any particular measurement method always read low by 1.1 ppm then you would not use the 1.1 ppm as an uncertainty. If you know it is 1.1 ppm low then you use that as a *correction* figure, not as an uncertainty. The uncertainty comes from the way the measurement is being made. Let's suppose you measure the voltage with a voltmeter and get one answer, then you measure it with a calibrator and null detector and get a different answer. This is quite usual. You will never get *exactly* the same answer using two different measurement methods. You then have to use considerable engineering judgement and expertise to decide what uncertainty should be attributed to the measurement. This unknown systematic error is the Type *B* uncertainty.

Let me give you a simple real example of this. I measured the standby current on a pc motherboard. The current was 130 mA measured on a moving coil meter [Avo 8]. Just to be sure, I measured it on a cheap DVM [Fluke 37] as well; 112 mA. I was more than a bit surprised by this since it was a simple DC current measurement, nothing very difficult. I immediately checked the calibration of both meters against a DC calibrator at 130 mA. Both read *exactly* 130 mA. Which was the correct reading of the pc motherboard current, given that they differed by around 15%? Personally I was more inclined to believe the Avo because it is an entirely passive device. The DVM on the other hand was being powered from a battery eliminator and is an inherently more complicated device.

When reading DC, the Avo should read correctly regardless of the waveshape, whilst the DMM may well clip on a very spiky waveform and therefore read lower on this measurement. In any case believing the higher value was safer in this application.

But the actual answer was much simpler than all this conjecture. The ***burden voltage*** on the Avo was 65 mV, whereas the burden voltage on the DMM was an amazing 690 mV. It is not surprising that the DMM read low when measuring on a 5 V circuit. Had the DMM been used on its 10 A range, its burden voltage would have dropped to more like 6.5 mV. [Yes, I got caught out on this one!]

Add the variances of all the Type *A* and Type *B* sources then square root the result to give the standard deviation. This is the uncertainty at a 1σ limit, assuming a Gaussian distribution (due to the ***Central Limit Theorem***). The 1σ limit does not give a very high confidence level for the measurement. It is therefore generally agreed to quote an *expanded uncertainty*, by taking the original uncertainty and multiplying it by *k*, the *coverage factor*. *k*=2 corresponds to a >95% confidence based on the Normal distribution function.

As an engineer one would prefer to err on the cautious side and quote larger uncertainties than experience might suggest. Unfortunately there is commercial pressure on national laboratories and accredited calibration laboratories to quote low uncertainties in order to be competitive with other laboratories. Remember that a calibration certificate quoting a lower uncertainty is worth more and therefore typically costs more.

Another term used on calibration certificates is *Test Uncertainty Ratio* (TUR). For "low accuracy" secondary standards it is easy to calibrate the equipment against something that is at least 10× more accurate. For example a 0.1% DVM can easily be calibrated against a 10 ppm calibrator. The calibrator is 100× more accurate than the DVM so the error in the calibrator can be neglected without any worry. The ratio between the accuracy of the unit being calibrated and the calibration source is the Test Uncertainty Ratio, a larger number being better.

Let's take a measurement example from a production environment. Suppose I am making 10.000K resistors with a spec of ±0.01% (= ±100 ppm). Suppose I am measuring them with a DMM which has a total accuracy equivalent to ±10 ppm. You might reasonably suppose that I would say that any resistors that measured better than ±90 ppm were good and could be shipped, whereas any that measured more than this were possibly bad and should either be thrown away, re-worked or sold as ±200ppm parts. This process is known as *guard-banding*.

In order to reduce costs, one "accepted practice" is to use a test uncertainty ratio of at least 4 and to set the test limit at the specification limit. This means that some faulty (out of tolerance) parts will be passed, and some non-faulty parts will be rejected.

In the literature on the subject, all sorts of mathematics have been played about with. The cost of throwing away 'good' parts has been evaluated against the cost of supplying faulty parts and various ratios between the costs have been proposed. This is an impossible situation because the manufacturer of a component or piece of equipment cannot judge the economic impact that an out of tolerance part will have on the customer.

I have very simple ideas on this subject. If I order parts and they don't meet the spec then I get very unhappy very quickly. I once had a batch of a crystal oscillator modules delivered to a new drawing. This was to a new tighter spec for which I was paying three times the price of the old parts. The spec called for 10 ppm accuracy over particular bands of time, temperature and supply voltage.

10 ppm is not accurate for a good frequency standard, but it is pretty tight for this type of non-adjustable commercial oscillator. I was able to measure them on the bench to 0.01 ppm accuracy using an *off-air standard* and a frequency counter. Two units out of the fifteen I measured were up to 10.6 ppm off of nominal. Needless to say the supplier got a phone call and the defective parts were returned.

The supplier then said that the parts were not out of spec. The test rig he measured them on was set to exactly nominal supply voltage and when run on the tester over and over the results showed a peak-to-peak variation of something like 1 ppm. At this point I was losing patience. I pointed out to the supplier that noise and calibration uncertainty on the test jig were his problem and not mine. When my spec says 10 ppm it meant 10 ppm and not 10 ppm ± noise ±calibration uncertainty. You see in a situation like this the supplier was at risk of losing *all* the business for an indefinite time. The cost would have been considerable.

The point I am trying to make is that there may be no acceptable cost associated with supplying demonstrably out of tolerance parts. If my measurement uncertainty was greater then theirs, I would not have been able to *prove* that the supplier was at fault. This is why I took the trouble to reduce my measurement uncertainty down to a minimal amount.

One could argue that it is wise to put a guard band around what the supplier quotes compared to what is actually needed. If I needed 10 ppm then maybe I should have asked for 8 ppm parts. Again this all comes down to cost. If you can do it then great; it gives more confidence that everything will be fine. But if you habitually use tighter tolerance parts than you need, you may be adding too much additional cost.

16.2 Measuring DC Voltage

To the ‘pure academic’, the term “DC voltage” may be offensive. The fact that the term DC stands for “direct current” means that terms such as DC voltage and AC voltage may seem a bit perverse; nevertheless such terms are very understandable and roll neatly off the tongue. As such they are widely used, even by experts. If you get “told off” for using terms like DC voltage then just recognise the argument being used and deal with it as you see fit. (DC current is even more offensive!) I make no apologies for using such terms.

DC voltage is probably the easiest quantity to measure, particularly from a standards point of view. When *transferring* voltages from calibrators to voltmeters the bandwidth involved is generally less than 50 Hz and if there is still too much noise, the bandwidth can be reduced by taking multiple readings and averaging them. The source impedances are low and therefore the capability of signals to capacitively interfere with the measurement is greatly reduced. However you should have realised by now that magnetic interference is coupled into loops, so the calibrator’s source impedance does not affect the induced voltage.

For the reasons stated above, DC voltage can be measured with greater accuracy than resistance or AC voltage. DC current, when it is measured as the voltage across a known resistance, can therefore be expected to be worse than both voltage and resistance combined. This does not apply at national metrology institutes, however, since they use *cryogenic current comparators*.

The measurement of the DC voltage from calibrators in order to calibrate a DVM is a specialist measurement done to get a defined uncertainty on the DVM itself. If this were to be the entire field of measurement then it would be pointless. There has to be a *reason* for doing all these inter-comparisons. There has to be a final *user* of the measurements. For example, a manufacturer of logic ICs tests his parts over the voltage range 4.70 V to 5.30 V. He puts a guaranteed operating range on his data sheet of say 4.75 V to 5.25 V. If your measurements are ultimately traceable to the same standards, you can calculate how accurately you are reproducing the manufacturer’s conditions. This is the key to making a measurement. Your readings are related to readings taken by somebody else at a remote location.

Obviously you need a device {meter} that is accurate enough for your purposes. In addition to this you must use an appropriate connection scheme and you must be able to estimate the uncertainty of the overall result. It is difficult to make broad quantitative statements about this area, but qualitative statements should suffice. It requires a certain amount of technical skill to make a successful measurement. The lower {tighter} the required measurement uncertainty, the greater the level of skill required.

If your measurement is done under all of the following conditions, you do not have to worry about the uncertainty in the measurement:

- The accuracy needed is not more than a few percent.
- The DVM accuracy is better than 0.1%.
- The source impedance is less than 20 kΩ.
- The DVM input resistance is greater than or equal to 10 MΩ.
- The common-mode noise on the source is less than a few volts.
- The ambient radiated fields are not higher than a few mV/m.
- The ambient electric fields and magnetic fields are low.
- The ambient temperature is 22°C ±6°C.
- There is no salt spray, coal dust, or other unusual environmental pollutants, or adverse environmental conditions.
- The resolution on the reading is not required to extend down below 1 mV.

That is actually quite a lot of things to consider in order to neglect the measurement accuracy! Otherwise you are going to have to consider all those factors when determining the overall uncertainty in your measurement.

As an example, if you measure the battery voltage on your car, then a simple hand-held DMM will give an acceptable reading for the purpose of seeing if your alternator {generator} is broken. You would probe the battery terminals with the engine idling and the electrical system loaded with say headlights and heated rear window on. If the voltage was less than around 13.2 V it would indicate a problem. [Mine measured 13.7 V on this test.] A 5% accuracy DMM would not have been suitable because a correct 13.7 V system could have read as 13.0 V, suggesting a defective alternator. The basic DMM uncertainty is the biggest unknown in this measurement situation [I used a 0.5% accuracy handheld DMM for this measurement.]

Now that was a pretty trivial measurement and yet the consequences of being wrong could be either buying a new alternator when it wasn't necessary, or failing to buy a new one when it was necessary and therefore breaking down [failing to start] because of a flat battery.

Realistically when you make a DC voltage measurement you will need to be considering all the factors in the bulleted list above, and either dismissing them or adding them into the overall measurement uncertainty. A skilled engineer would not even notice that he was doing such a process. It would be 'obvious', 'common-sense' or 'unconscious'. So, in order for you to approach the same level of skill as an experienced engineer, without waiting for 20 years for you to get experienced, I have written these out for you as a checklist to follow. Once you get the idea, you won't need or want to follow it!

DC Voltage Measurement Accuracy Checklist

❑ Is the DVM uncertainty at least 4× lower than the required uncertainty?

❑ Estimate the source resistance, R_S . Is DVM input resistance $>10\times R_S$?

❑ Source loading error is roughly, $100\% \times \frac{\text{source resistance}}{\text{DVM input resistance}}$.

❑ If the source resistance is greater than 10 kΩ, have leakage current errors been minimised?

❑ Is the DVM noise significantly higher than when the DVM inputs are shorted? If so, check for mobile phones, soldering irons, nearby faulty equipment &c.

❑ Is there noise on the DVM that can be reduced by twisting or shielding the leads?

❑ Does the reading look noisy, suggesting the DVM is causing an oscillation? (Use a stopper resistor in series with the Hi lead at the circuit end of the lead.)

❑ Is the capacitance of the DVM and test leads causing the circuit to oscillate? (check using a scope or spectrum analyser at some other point in the circuit).

❑ Can filtering or averaging be used to reduce the measurement noise?

❑ Is the mean value still drifting because insufficient time has been given for the equipment and/or the reading to stabilise?

❑ Is the ambient temperature correct for the DVM or has an allowance for the DVMs TC been made?

❑ Is the DVM guard box (or the Lo terminal if there is no explicit guard) connected to a low impedance source of common-mode for the measured system?

❑ Is the magnitude of the reading significantly different when the DVM connections are reversed?

❑ If the measurement resolution is below 200 μV, have thermal EMFs been minimised and allowed to stabilise?

❑ For uncertainties below 20 ppm, note temperature & humidity.

❑ Is the lighting level affecting the measured value?

❑ Is strong sunlight directly hitting the system, causing rapid heat fluctuations with cloud movements?

❑ Evaluate the overall uncertainty using a linear sum or an RSS combination.

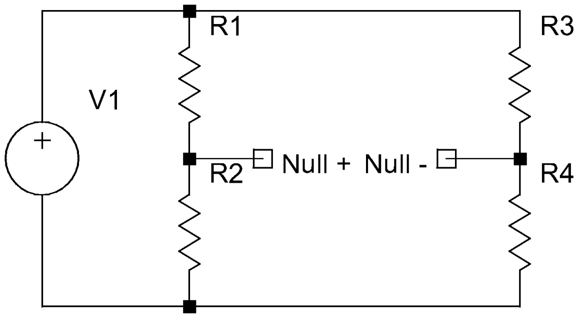
16.3 Measuring Resistance

Resistance is probably the second easiest electrical quantity to measure, depending on its value. It is certainly difficult to measure resistors below an ohm with great accuracy, and it is also difficult to measure resistors above 10 MΩ with great accuracy. The thing is to specify just exactly what is meant by "great accuracy". For some engineers 0.1% is great accuracy, whereas for others 0.05 ppm would be great accuracy; it depends where you are and what you are trying to do. It is not unusual to want to measure resistances < 1 μΩ and it is equally not unusual to want to measure resistances > 1 TΩ (that's $>10^{12}$ Ω). Clearly the measurement techniques for these resistor values are quite different to each other, the range being in excess of 18 orders of magnitude!

The Wheatstone-Christie Bridge

Devised by S.H. Christie in 1833,[1] this bridge was popularised by Sir Charles Wheatstone of King's College [London] in 1843. It has been referred to as the *Wheatstone Bridge* for many years, but the more appropriate name *Wheatstone-Christie Bridge* will be used here. Its purpose was to measure resistance "accurately" and many configurations are considered as variants of this basic type.

FIGURE 16.3A:



In this scheme the voltage source does not need to be accurate or even stable. By having three known resistances, the fourth can be determined. At least one of the resistors is adjusted so the detector across the centre of the bridge reads zero (a null condition).

Various ingenious schemes, using switched resistors in the *arms* of the bridge {each of the resistor positions is an arm in the bridge}, have been arranged to make useful measuring boxes. Today these schemes are only seen in elementary physics classes, if at all.

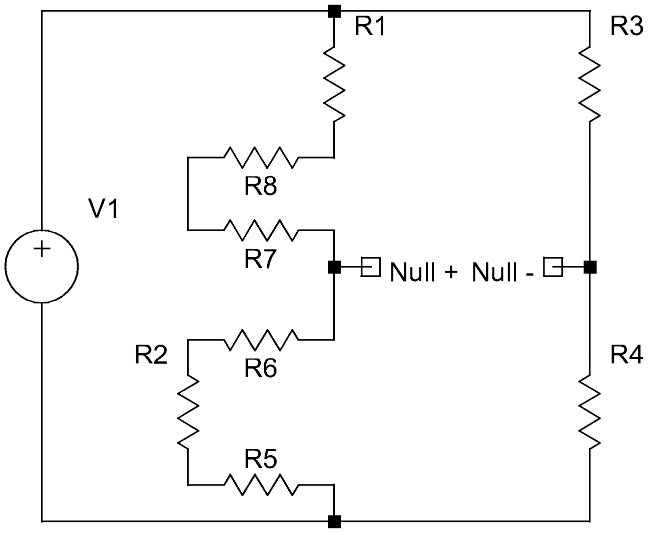
These bridges are often drawn in the shape of a square rotated 45°.This depiction is meant to show the lack of resistance between the ends of R1 and R3, but most circuit drawing packages don't like components that are anything other than vertical or horizontal.

If the unknown resistor is on the left side of the bridge, the right hand side could be a potentiometer. Wheatstone found it easier to use a long single-wire 'potentiometer' to form the R3-R4 part of the bridge. A piece of high resistance wire is laid against a metre rule, allowing the position of the null to be read off the scale. It should be easy to measure the distance of the null point to within 1 mm, giving a resolution of 0.1%. That is very old technology, so there is no point in looking into the ways of improving that old equipment.

It is not practical to plug all types of resistor straight into the measuring box; it would be more convenient to measure the resistor at the end of some leads.

[1] J.C. Maxwell, Index footnote, in *A Treatise on Electricity and Magnetism*, 3rd edn (Clarendon Press, 1891; repr. Dover Publications, 1954), Vol 2.

FIGURE 16.3B:



Nominally equal lead resistances R5, R6, R7 and R8 have been added to the circuit diagram. Suppose R2 is the resistor to be measured. R5 and R6 are the *go* and *return* wires to R2. The 'dummy' wires R7 and R8 are supposed to be equal in length to R5 and R6. They go and return without connecting to anything else. If R1 is close in value to R2 then the lead resistance error has been removed to some degree.

If R4 has a similar value to R2, the lead compensation resistors could be put in series with R4 instead of R1. Lead compensation is done on strain-gauge bridges and similar applications. This technique is used for long lead lengths, but it is a *compensation* scheme and clearly relies on the matching of the lead resistances and the joint resistances. Its accuracy is therefore limited, particularly when the resistor to be measured has a value similar to, or lower than, the lead resistance.

Some people think that this lead-compensation bridge scheme is what is meant by a *4-wire resistance measurement*; it is not, despite using 4 physical wires to measure the resistance.

2-wire ohms measurement

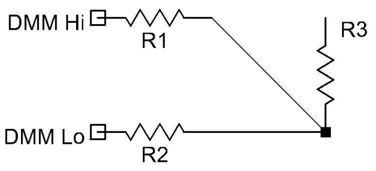
Consider using a moving coil ohm-meter. You zero out the lead wires before you start, hook up the resistor, then measure its value. But if you look at the resistance scale on your moving coil ohm-meter, you will see that it starts off easy to read on the right hand side of the scale, but it gets more and more compressed as it heads to the left. This is a *non-linear ohm-meter*, meaning the scale is non-linear. When digital ohm-meters first appeared the banner headlines were *linear ohm-meter*, the latest greatest invention!

EX 16.3.1: A moving coil ohm-meter consists of a 50 μA FSD meter in series with an adjustable resistor and a 1.5 V battery. It is trimmed to read full scale when there is 0 Ω external resistance in the circuit. [Keep it simple. Leave the battery at 1.5 V and the meter movement as exactly 50 μA &c.]

a) What is the nominal value of the internal resistor plus the meter movement?
b) At what positions on the scale should the markings be for resistances of 10 Ω, 100 Ω, 1 kΩ, 10 kΩ, 100 kΩ & 1 MΩ?

Having seen the non-linearity problem with the moving coil ohm-meter, let's move to a high-accuracy linear digital ohm-meter.

FIGURE 16.3C:



The first vital thing to do is to *zero* the meter. Suppose the meter leads, represented by R1 and R2, are 1 m long. It is no good putting a shorting link across the terminals of the DMM to get the zero. The zero is obtained by shorting the leads together at the resistor to be measured. This is pictorially represented in the circuit above. Now, when you move the HI wire to the top of R3 you get "no error" due to the leads.

There is an error, but if you are measuring a 10 kΩ resistor with a 3½ digit DMM you won't see it. If you are measuring a 1 Ω resistor with a 6½ digit DMM the resolution is 1 μΩ. Now you will get an error.

***EX 16.3.2:** List the sources of possible error in the 2-wire resistance measurement scheme just due to the inter-connection scheme.

We need a number for variation of lead resistance and contact resistance; I am going to pick 10 mΩ. You could argue with me on that, and perhaps use a value 5× larger or smaller. The problem is *you can't measure this value in the measurement situation*. You can test connections by making them and unmaking them several times. This gives you a variation and from this you can estimate the worst possible error. I tried making connections with a gold-plated spade connector in several positions on the same terminal post and I could easily convince myself that the variation was around the 1 mΩ level. But I have to account for any set of leads, and any set of terminals, so the uncertainty must be larger.

RESISTOR BEING MEASURED	ESTIMATED 2 WIRE MEASUREMENT UNCERTAINTY
1 mΩ	☹
10 mΩ	☹
100 mΩ	10%
1 Ω	1%
10 Ω	0.1%
100 Ω	100ppm
1 kΩ	10ppm
10 kΩ	1ppm
100 kΩ	0.1ppm
1 MΩ	0.01ppm
10 MΩ	0.001ppm+

It is up to you to establish what change of resistance is possible in *your* connection scheme and then either neglect it, take it as an uncertainty, or wire the circuit up better to minimise that error.

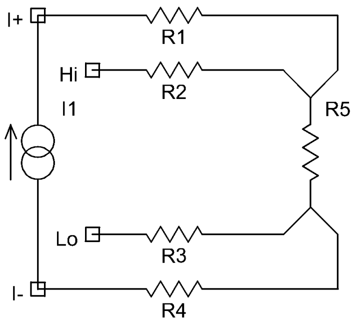
You should now see why laboratory standards resistors up to 10 kΩ are all 4-terminal devices.

4-wire ohms measurements

The terms *4-terminal measurement* and *4-wire measurement* are interchangeable. Whilst I am going to show you a circuit representing a 4-wire ohms measurement, please don't get the idea that you have to be measuring special 4-terminal resistors to make use of this technique. You can still get an improved measurement of a two terminal resistor by using this 4-wire technique.

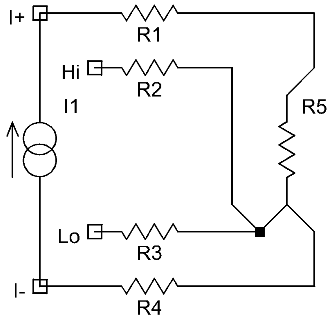
I am assuming that you have a 4-wire DMM to do this measurement with. If you don't, you can still construct a measurement based on the same sort of principles using precision calibrators and a null detector.

FIGURE 16.3D:



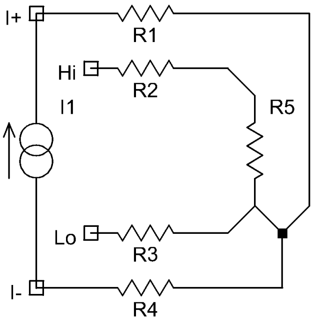
The current source, I1, in the DMM supplies the measured resistor via the lead resistances R1 & R4. The voltage across the measured resistor R5 is sensed with a high input resistance DVM connected between Hi and Lo. The finite output resistance of the current source and the finite input resistance of the DVM section limit the overall rejection {reduction} of the lead resistance. The first part of the measurement is to do a *4-wire zero*.

FIGURE 16.3E:



This is one way of doing the zero for a 4-wire ohms measurement. The resistor runs at the same current during the zero and the measurement, maintaining the operating temperature; thermal gradients are therefore unchanged during the zeroing operation. This method changes the source impedance to the DVM section, giving an offset due to the bias current of the DVM section.

FIGURE 16.3F:



This is another way of doing a 4-wire zero. Using this connection scheme, then moving the Hi connection lead down to the Lo terminal will establish the magnitude of the bias current offset created by the DVM section.

You may get different answers between these connection schemes and it is necessary to evaluate the measurement theoretically to see which is the most accurate. With this subject there is no ‘correct answer’; you cannot change the connections around until this *correct* value is achieved. It is important to try both methods to see what size of difference is caused by the connection scheme alone. This will help to estimate the uncertainty in the measurement.

Although moving the Hi sense lead keeps the resistor at the same temperature, the benefit of this is lost to a certain degree because the sense lead is moved away from this temperature gradient. On the other hand, moving the I+ force lead definitely changes the power distribution in the system.

In general, a connection method that gives less noise and better repeatability is more likely to give an accurate reading of the desired quantity. Another general point is that thermal EMFs in the force leads, I+ and I–, are entirely irrelevant. However, thermal EMFs in the sense leads, Hi and Lo, are very important.

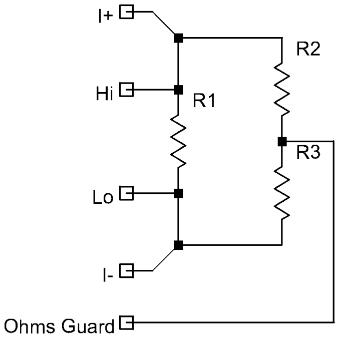
The key point to note is that the DMM will be most accurate when used according to the way it was calibrated. If it was calibrated using a zero as described in the first circuit, then it is best to do your measurement in the same way. The manufacturer should explain how they do the 4-wire zero in the manual.

The 4-wire connection scheme for resistance measurement is replaced by a *four* *terminal pair* (4TP) scheme when measuring impedance. The four wires are replaced by 4 coaxial cables in order to deal with the shielding problems associated with the AC measurement.[2] The coaxial screens more accurately define the impedance because stray reactive effects are then reproducible, reducing the uncertainty between different measurement labs.

Ohms Guard:

An Ohms Guard is not related to shielding from electrostatic interference, it is related to *leakage paths*. These could be due to making an in-circuit measurement, or could be related to possible leakage paths across a high value resistor. The equivalent circuits are the same, and the guarding technique is the same, so they come under the same heading.

FIGURE 16.3G:



This circuit demonstrates a 4-wire guarded resistance measurement of R1, but if this resistor has a high value (>10 kΩ), a 2-wire plus guard measurement might be acceptable.

The in-circuit shunt resistance is represented by the resistors R2 and R3. [If a physical resistor is shunted directly across R1 then it cannot be eliminated by this technique.] R2 and R3 could also represent a leakage path.

If the Ohm's Guard is held at the same potential as the Lo terminal there will be 'none' of the shunt path current flowing into the I– current measurement terminal. There will be excess current flow in the I+ path, but the measuring device would measure only the I– current. The leakage path is therefore eliminated, or at least minimised.

***EX 16.3.3:** R1= 10 MΩ; R2= 900 GΩ; R3= 100 GΩ. Neglecting the error due to the meter, the leads, temperature and everything other than shunt resistances:

a) What is the measurement error if the guard isn't used?

b) What is the measurement error if the guard (referenced to Lo) has a ±10 mV offset, the measurement being done at 1 V?

It should now be evident that high value precise resistors need to have a guard terminal *on the component*. The guard offset error given above is excessive and would normally be expected to be <100 µV. Nevertheless, any sort of guard gives a much improved accuracy.

The extreme of ohms guard is found on in-circuit component testers. In this situation the shunting resistor is allowed to be anywhere from 1000× to 1,000,000× *lower* than the resistance being measured. The measurement accuracy is reduced to the order of magnitude of ±1% for this type of measurement however.

Whilst the surface of a component will undoubtedly be sufficiently insulating when the component is new, after years of use it will have become caked {coated; covered} in dust and environmental pollutants, significantly degrading the insulation performance.

[2] B.P. Kibble, and G.H. Rayner, *Coaxial AC Bridges* (Bristol, UK: Adam Hilger, 1984).

With modern high quality materials there should not normally be a problem with the solid material of the insulators, but there can still be a problem from surface contamination. Older materials would have particularly suffered from moisture absorption and therefore relative humidity would have potentially been a more serious issue.

1956 technology → “If the relative humidity exceeds 65 percent, a conducting film of moisture will be adsorbed on the surface of many types of insulating material. Leakage over such surfaces, particularly to the circuit of a sensitive galvanometer, may introduce serious errors.” [3]

The only way to prove the point is by measurement. How much does relative humidity change the value of the actual component you are measuring?

This next table gives an idea of the order of magnitude of shunt resistance that can cause measurement problems.

RESISTOR SHUNTED	ERROR DUE TO SHUNT RESISTANCE					
	100 MΩ	1 GΩ	10 GΩ	100 GΩ	1 TΩ	10 TΩ
10 kΩ	100 ppm	10 ppm	1 ppm	0.1 ppm	0.01 ppm	0.001 ppm
100 kΩ	0.1%	100 ppm	10 ppm	1 ppm	0.1 ppm	0.01 ppm
1 MΩ	1%	0.1%	100 ppm	10 ppm	1 ppm	0.1 ppm
10 MΩ	☹	1%	0.1%	100 ppm	10 ppm	1 ppm
100 MΩ	☹	☹	1%	0.1%	100 ppm	10 ppm

I found an old laboratory standard 1 MΩ resistor that had been taken out of service and was covered with a reasonable dust film. Keeping the DMM connections in place, I wiped the dust off and the DMM registered an increase of something like 220 ppm for a few tens of seconds before returning to its previous value. Evidently there was some static electricity built up by rubbing the plastic top with a cloth. The conclusion is that for terminals that are 80 mm apart, it is acceptable to neglect contamination films on the surface in terms of the measurement accuracy up to the level of a few ppm. In this case the contamination film must have been greater than 100 GΩ (according to the table).

This next exercise relates to in-circuit measurements.

EX 16.3.4: Using the same circuit model as before, R1= 1 kΩ; R2= 100 Ω; R3= 10 Ω. Neglecting everything else, what is the error in the measured value of R1 if the ohm’s guard circuit wire has a resistance of 10 mΩ?

10 mΩ of lead resistance is a very reasonable figure, so in-circuit measurements like this can be quite inaccurate. The error can be reduced by using a *guard force* and a *guard sense*. This two-wire guard approach minimises the guard wire resistance error for in-circuit measurements.

Measuring Resistances above 10 MΩ

For simplicity I won’t keep mentioning and drawing the ohms guard, but all measurements in this section require the ohms guard to remove any leakage paths.

[3] F.B. Silsbee, 'Suggested Practices for Electrical Standardizing Laboratories', *Laboratory Atmosphere*, [also in NBS Handbook 77-vol I], *National Bureau of Standards Circular 578* (Aug 1956), page 2.

I am making the assumption that your measuring device does not measure above 10 MΩ, but can measure to 10 MΩ. Obviously there are instruments specifically designed to measure high resistances; if you are doing a lot of high resistance measurements then you should get one. As an example, electricians use high voltage resistance testers to check newly installed or altered wiring. These meters can typically measure up to at least 1 GΩ with a test voltage of up to 1000 V DC.

There are two specific reasons for measuring high values of resistance. One is for ultra-low leakage circuits and the other is for high voltage work. However, it is not acceptable to measure a high voltage circuit's leakage resistance at a low voltage.

High voltage circuits, and that could mean anything above 100 V really, are not guaranteed to behave linearly with applied voltage. They arc-over for example. For this type of circuit use the second of the methods I am about to describe.

The first method is for low voltage circuits and is a safe, but low accuracy method of measuring high resistances. Suppose the DMM measures to 10 MΩ. Set the DMM on the 10 MΩ range, do an 'input zero', then connect a high quality 10 MΩ resistor directly across the input terminals, leaving it to settle down for a few minutes. This will give you an idea of the noise on the reading and any temperature drifts that are occurring.

Leaving the 10 MΩ resistor in place, connect your test leads, leaving the measurement end open-circuit. Don't let the metal conductors touch *anything*. The wires can be bent so that they stick up from the bench. They can be hung over the bench. They can be propped up out of the way with erasers, reels of sticky tape, staplers, or anything else lying around. They just must not touch anything else. Now, look at the displayed reading on the DMM. If it has changed by more than a few digits in excess of the noise that you first observed, you may have a problem with leakage currents in the leads. Check by repeating the initial test to see if that reading has drifted.

Now hold the insulating part of the leads in your hands to see if the reading changes. This tests whether or not the DMM is capable of rejecting the noise {hum} induced by your body picking up ambient fields and injecting them into the measurement. It is quite likely that you will not be able to hold the leads without changing the reading. You should also try moving around when you are not holding the leads to see if your body position affects the reading, and if so by how much.

You have now established the negligible or non-negligible effect of the leads in this test set-up, at this resolution. Now you can clip or connect the leads to the resistor you want to measure and take a new reading, after allowing for settling time. Watch the display for a minute or so and see how fast the digits are changing. Leave it until the digits aren't changing due to drift (they may run around due to noise, but you will notice this as well).

During this settling process *do not move around*. Your body may affect the readings. You should ideally be in the same position as when the resistor across the DMM terminals was being measured. Do not create additional uncertainty in the reading by something as silly as standing up during one set of readings and sitting during the next.

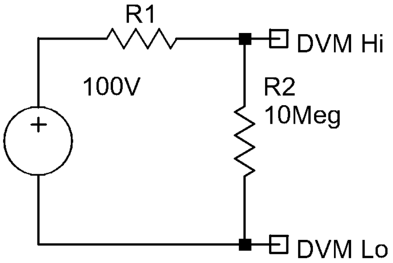
The additional shunt resistance can be calculated from the two readings.

***EX 16.3.5:** You have a 5½ digit DMM with a spec on the 10 MΩ range of ±500 ppm of reading ±2 digits. You zero the meter on the 10 MΩ range, then you place a good quality 0.1% 10 MΩ resistor across the input terminals; it measures 10.0016 MΩ. The ambient temperature is stable for the duration of the test. The meter is running well inside its specified conditions. The reading on the display is very stable and it just sits there mostly reading the figure shown above. Occasionally it changes to 10.0015. You connect the leads up and there is no change in the value. Possibly it reads 10.0015 slightly more often than it used to. You apply the resistor to the leads and the reading drops to 9.9903 MΩ.

a) What is the nominal value of the unknown resistor?
b) What is the measurement uncertainty of that value?

Now you can see that you can measure a 10 GΩ resistor without much difficulty, but only to a few percent accuracy. It looks from this as though you should be able to measure a 100 GΩ resistor, but with further reduced accuracy.

FIGURE 16.3H:



This is the second of the methods of measuring high value resistors, but this one is done at a more suitable voltage such as 100 V or more. In this arrangement you connect the circuit up with the voltage source set to 0 V. When the DVM reading stabilises you set the supply to 100 V and wait for the reading to settle again. The settling time is important because any dielectric present will allow some current flow at first. This initial lower resistance reading is due to *dielectric absorption* in the capacitance of the measured system. This dielectric effect is particularly important for unintentional resistances, such as leakage paths from a relay's coil to contact.

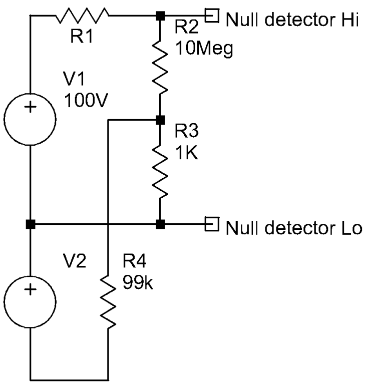
This technique works best with one of the accurate >10 GΩ input resistance DVMs described earlier. However, if you only have a 4½ digit DVM with 10 MΩ input resistance, you can still use this method; the 10 MΩ resistor is just omitted.

***EX 16.3.6:** You wire up the circuit shown for measuring high value resistances using the high voltage source. Your DVM has 5½ digits, >10 GΩ input resistance, <50 pA bias current. Its lowest range is 100 mV. When the voltage source is at 0 V the DVM reads 0.325 mV. When the voltage source is at 100 V the DVM reads 1.259 mV. The DVM has an averaging feature and you are using it, and the input filter, so the reading is stable after you leave it for a few minutes. The voltage source has been measured on the DVM and you can say that it is 100 V ±0.1%. The 10 MΩ resistor hasn't been measured recently, so is only known to be ±0.1%. The DVM spec is 50 ppm of reading ±1 digit ±1 μV. The DVM is warmed up, the temperature is not drifting, and there is no external temperature gradient.

a) What is the nominal value of the unknown resistance?
b) What is the uncertainty of the measurement?

How much better could I do if I tried harder? I could measure the voltage source and the resistor to 100 ppm accuracy instead of 0.1% (1000 ppm), but that doesn't increase the overall accuracy by a factor of ten because the dominant effect is the linearity error in the DVM.

FIGURE 16.3i:



Suppose the null detector has 100 nV resolution without appreciable noise, a realistic expectation. Resistors R3 and R4 are ratio-matched for a stable 100:1 division, which can be characterised. V2 is used to null the system with V1 set to 0V. With V1 then set to 100 V, the null is obtained again using V2. The null detector voltage is constant between the two readings and therefore so is its input current. With R1 around 1 TΩ the voltage across R2 will rise to about 1 mV, requiring 99 mV across R4. The power in R4 is 0.1 μW, giving <20μ°C even at 200°C/W. The self-heating effect in R4 is therefore negligible. Each null is accurate to 100 nV in 1 mV, representing 100 ppm uncertainty. If you also measure V1, V2, R2 and R3/R4 to an accuracy of better than 100 ppm each, the resulting measurement uncertainty is less than ±600 ppm (±0.06%) even using a linear sum of uncertainties. A further factor of ten improvement in all the measured values is also a practical proposition.

For a voltage measuring device such as a DVM, or for a pass-thru device such as an amplifier or filter, the input resistance under actual working conditions is easy to measure. Apply a signal from a low impedance source, then apply the same signal through a large value resistor. The drop in signal level on the display, or measured at the output of the amplifier/filter, will allow the input resistance to be calculated.

***EX 16.3.7:** An instrumentation amplifier is driven directly from an AC calibrator at 100 Hz. The amplifier output is fed into a true RMS DVM set on its AC range. The calibrator output is adjusted to make the DVM read approximately 10.0000 V on average, although the last two digits are ‘running around’ quite a lot. The amplifier input is then fed via a wire-ended 10 MΩ resistor, the resulting DVM reading being 9.102 V approximately. Estimate the amplifier’s input resistance and comment on the technique.

A DVM input resistance can be measured in a similar way but using a DC calibrator. Use 10 MΩ // 10 nF in series with the Hi lead to the calibrator. The Lo lead goes directly to the calibrator. Set the calibrator on 0 V and see the effect of shorting out the resistor//capacitor pair. The change in reading is the bias current error. Using the 10 V range on the DVM, apply 10 V from the calibrator. The change in reading when the resistor is shorted represents the new bias current at this elevated input level. The 10 V input swing divided by the change of bias current is the input resistance.

Measuring resistances below 10 Ω

Measuring resistance at 100 mV		
RESISTANCE	CURRENT	POWER
10 Ω	10 mA	1 mW
1 Ω	100 mA	10 mW
100 mΩ	1 A	100 mW
10 mΩ	10 A	1 W
1 mΩ	100 A	10 W

Low value resistances are best measured at high current, in the same way that high value resistances are best measured at high voltage. One difficulty can be that the test current causes too much self heating. Suppose you want to measure at a voltage of ≥100 mV in order to avoid resolution and thermal EMF problems.

Using a smaller test voltage reduces the available resolution, reducing the measurement accuracy. Another point is that measuring a resistor accurately when its value is lower than the interconnection leads obviously requires excellent rejection of the lead resistance; 4-wire measurements are therefore essential.

Measuring at higher power is not necessarily a problem however. If the application is as a current sense resistor for use at 10 A then you will get the greatest accuracy *for the intended purpose* if you measure the resistor under the same conditions as those in which it will be used. This includes, but is not limited to, operating current, operating ambient temperature, humidity, and mechanical stress.

Let's suppose you are using the resistor as a 10 A current sense resistor and you measure it at 1 A. There is then a power coefficient uncertainty to add to the measured value before it can be used. Thus you can minimise the *end-user tolerance* {uncertainty} by adjustment of the measurement method.

The main measurement problems you are facing are lead connections and lead rejection, power dissipation, and thermal EMFs. The problem with thermal EMFs is actually relatively easy to deal with. Because of the self-heating of the resistor, and some notional internal or external thermal asymmetry, one end of the resistor will be hotter than the other end. Thus a thermal EMF will be generated with respect to the measurement system. The direction of this thermal EMF will be substantially independent of the direction of the current. Thus reversing the current direction, subtracting the voltage readings and dividing by two will give the true voltage reading, provided the DVM itself has a relatively negligible error on positive/negative reversal error. Reversing the current direction then gives a more accurate measurement of the resistance.

$$V_1 = \; I_S R_{UNKNOWN} + V_{THERMAL}$$

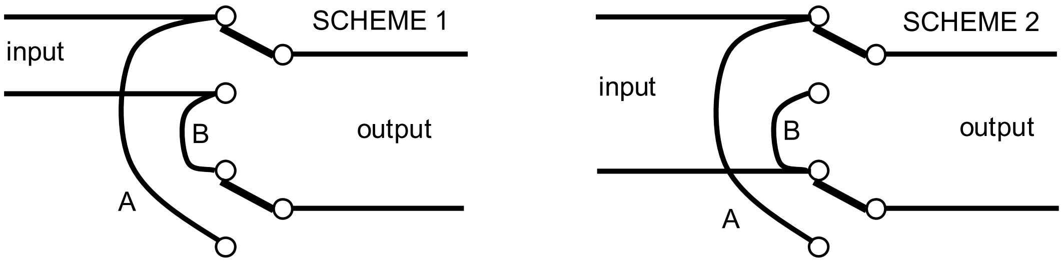
$$V_2 = -I_S R_{UNKNOWN} + V_{THERMAL}$$

$$\frac{V_1 - V_2}{2} = I_S R_{UNKNOWN}$$

This is an easy method to use with a special purpose switching matrix. You also find this technique used in commercially available equipment.

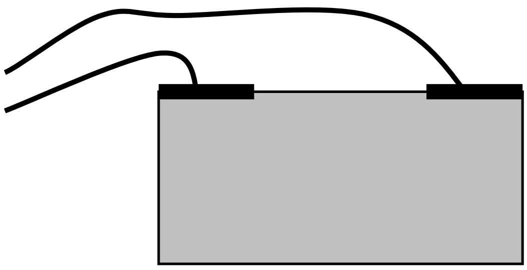
EX 16.3.8: What errors might be introduced by this method?

***EX 16.3.9:** The diagram below shows two subtly different ways of physically wiring up a reversing switch. When used on the force leads of a 4-wire precision measurement system running at 100 A, is one scheme better than the other? (Hint: consider the resistance of wires A and B.)



The next type of error, which is even more of a problem with low value resistances, relates to the idea of 'current spreading'.

FIGURE 16.3J:



This diagram represents a planar resistive film with two electrodes connected in an unhelpful manner as far as measuring the resistance of the film is concerned. It is a matter of 'common-sense' to an electrical engineer that there would be a massive *fringing field* in this situation.

Such two-dimensional field distributions are usually analytically intractable. Some are solved by symmetry, some require *conformal mapping*, but all are very straightforward to solve using computer 2D field solvers.†

Now a two-dimensional field is much easier to visualise than a three-dimensional field, but the principle is exactly the same. The current has to spread out over a finite distance until it is reasonably uniform through the conductor. Any intelligent person would immediately resolve this problem by re-positioning the electrodes to the ends.

You could now attach voltage measurement probes of small dimensions at points near to the end caps and you would get a reasonable measure of the resistance of the rod. However, what happens when the rod is made of high conductivity material such as copper?

The obvious first answer would be to use higher conductivity end plates, making the effect of current spreading in the end plate 'insignificant'. The table shows that using a lower resistivity material is not a feasible solution, since no such conductor exists! All you can do is make the resistance definition point a larger distance away from the current feed points. This means making the conductor longer to give the current a chance to spread out.

metal	Resistivity (μΩ·mm)
Silver	16
Copper	17
Gold	24
Chromium	26
Aluminium	27

If the resistance is always measured at the same points then the fringing effect should be constant. The reason why it gives measurement uncertainty and non-reproducibility is that the connection to the current feed points will be variable from measurement to measurement and therefore the spreading pattern will be different. You will consequently get better measurements by reducing dependency on this spreading current distribution.

You should now understand that just using *Kelvin clips* does not guarantee a successful, repeatable and accurate measurement. Certainly you will get a better answer than by using a two-wire measurement, or by making a 4-wire measurement only up to the point of making contact with the measured component. Just realise that you need to understand what you are doing in order to get a result with a low uncertainty.

Understanding that you have to take all the previous points into account, I will not explicitly be mentioning them in this next section. The subject is now more about the measurement equipment. I have been stressing the use of a 4-wire system as opposed to

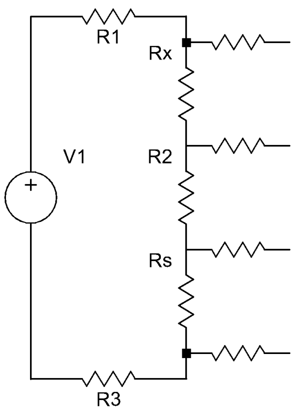
† Try 2DField from **www.logbook.freeserve.co.uk**

a bridge measurement. It must be noted, however, that there is a bridge method which is also a true 4-wire method. The system is known as the *Kelvin double bridge*. It was devised in 1861 to overcome the problem of measuring the resistance of metal 'bars' 6 mm long and of 1 mm square section.[4]

EX 16.3.10: What is the approximate resistance of these 'bars' if made of copper.

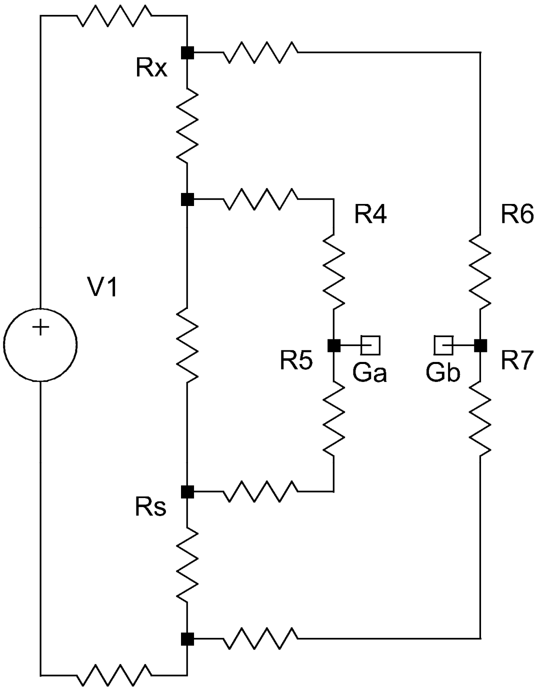
These copper bars have such a low resistance that the task is not trivial even with modern technology. I am not going to address the method of connecting to the rods. You obviously need 4-wires with the sense wires some distance in from the ends. This requires considerable mechanical ingenuity. What I am going to explain is the electrical measurement method.

FIGURE 16.3K:



The first step in the development of the double bridge is to draw out the main current path. The standard resistor is Rs and the unknown is Rx. R1 and R3 are the lead resistances to the battery and include the connection resistances to the unknown. R2 is the lead resistance between the standard and measured resistance and it includes the connection resistance to each of the resistors. Because low value resistances are measured using this bridge method, R1, R2 and R3 are not negligible. In fact they may be larger than Rx and Rs. The unmarked sense resistances may also be larger than the measured resistors.

FIGURE 16.3L:



This is the full Kelvin double bridge. The unmarked resistors are the lead and contact resistances whose effects you are trying to minimise.

Suppose the lead and contact resistance on the sense wires are around 100 mΩ. Making R4, R5, R6 and R7 large compared to this makes the contact resistance error arbitrarily low. Make the leads in pairs so that the pair connected to R4 and R5 are as equal as possible; likewise for the other pair of sense leads. The accuracy of the measurement is reduced down to the accuracy of the ratio pairs R4/R5 and R6/R7. For each of these pairs, the wires can be swapped over to reverse the direction of the ratio error. Hence by making 4 measurements the ratio error of the external resistors can also be minimised.

It is not essential for the ratio pairs to be exactly equal, but the accuracy is greatest

[4] W. Thompson (Later to become Lord Kelvin), 'On the Measurement of Electric Resistance', in *Transactions of the Royal Society*, XI (1861), pp. 313-322+.

when they are equal. There are a great many possible subtle variations of this basic concept, but once you understand this fundamental circuit, the rest are easy to follow.

Just as a reminder, don't forget that thermal EMFs, power balance, thermal gradients, contact pressure, current spreading, current reversal, and shielding are all important in these measurements. Use twisted pairs for the sense leads just out of habit, even if you think that radiated pickup is not problem.

Measuring Impedance

Whilst resistances are typically measured with DC, impedances must be measured with AC. The 4 wire measurement is now replaced by the *4-terminal pair* measurement, using four coaxial connections. The screened coaxial connections minimise strays and allow capacitance measurement to resolutions below 0.1 fF.[†]

16.4 Measuring AC Voltage

To accurately measure AC voltage in a standards laboratory, it is usual to have an accurate DC voltage available and to use a *transfer device* having a well defined characteristic comparing its response to AC and DC voltages. A common device for this purpose at low frequencies (1 MHz) would be a *thermal transfer standard*. The heating effect of a DC source is compared to the heating effect from an AC source. Clearly this is doing an RMS comparison of the AC source against the RMS value of the DC source. The DC source needs be relatively noise free {quiet} in order to make the RMS value close to the mean {average} value. For RF, microwave and mm-wave usage the AC/DC transfer device would be either a micro-calorimeter or a bolometer.

Since it is required that the thermal device should average out the incoming power fluctuations, the thermal transfer is not as accurate at frequencies below a few tens of hertz. Typically the best transfer accuracy would therefore be achieved in the 200 Hz to 20 kHz region, with better than 20 ppm uncertainty.

The requirement of low noise on the DC source can be tested as follows. If the DC source is considered to be a pure DC value with a small AC amount (of normalised RMS value δ), then the mean value will be just the DC amount, but the RMS value will be $\sqrt{1+\delta^2} \approx 1+\frac{\delta^2}{2}$. The noise requirement is evidently not difficult to achieve for errors of the order of a few ppm.

δ	RMS NOISE	$\Delta\left(\frac{RMS\,\mathrm{DC}}{mean\,\mathrm{DC}}\right)$
0.1	10%	0.5%
0.01	1%	50 ppm
0.003	0.3%	5 ppm
0.001	0.1%	0.5 ppm

Below 30 kHz, AC signals can be routed via ordinary wires without giving rise to gross measurement problems. Above 1 MHz, the routing of AC signals is usually done with coaxial cables and connectors.

A 1 metre length of cable or leads might have a self-capacitance of between 30 pF and 100 pF. Taking the high value, it is clear that at 100 kHz the shunting effect of the cable is $Z=1/(2\cdot\pi\cdot100\times10^{-12}\times1\times10^{5})=16{,}000\ \Omega$. This is not an effect that should be ignored. If the capacitance varies then the loading effect will also vary, so it is sensible to only use leads with defined separation between the conductors. Coaxial cable meets this need and provides shielding as well. The shielding is *reciprocal* in the sense that the

[†] Agilent E4980A precision LCR meter, for example.

cable does not radiate whatever signal is flowing through it and also it does not pickup external fields.

Whilst DVMs are available that measure up to 1 MHz and beyond, their use is strictly limited to measuring sources with very low output impedance at these frequencies. Otherwise, the cable and voltmeter loading effect will not be small. However a low source impedance driving an open-circuit or dominantly capacitive load gives rise to a nasty resonance effect. Well away from this resonant point the per-unit voltage error can approximated by $\boxed{\Delta \approx (2\pi f)^2 t_d \left(CZ_0 + \frac{t_d}{2} \right)}$, where C is the capacitive load, t_d is the propagation delay down the cable whose characteristic impedance is Z_0. This approximation is only valid when Δ remains below 0.05 (5% error).

EX 16.4.1: An AC voltmeter has a 1 MΩ//30 pF input impedance. It is connected to a source having a resistive output impedance of 100 Ω via a cable of 100 pF. Treat the cable as 'short' and lossless. Treat the input and output impedances as exact.

a) What is the loading error at 100 kHz?
b) What is the loading error at 1 MHz?
c) If the source has 0 Ω output impedance and the cable is 5 m of 50 Ω coax, how much peaking will result at 1 MHz, assuming a dielectric constant of 4?

Get used to the fact that AC measurement uncertainties are much larger than DC measurement uncertainties. Whilst DC voltages can be measured to uncertainties below 1 ppm, typical AC voltage uncertainties are more like 10 ppm at 1 kHz and 1% at 1 MHz.

As the frequency increases there comes a point where everything is done from 50 Ω sources, is connected via 50 Ω coax, and is measured using equipment with 50 Ω input impedance. Working out what uncertainties are involved in an AC measurement then becomes remarkably involved.

A whole new set of theory is required to work with 50 Ω systems or waveguide, but this theory is usually only taught on optional microwave or RF courses. The subject areas involve transmission lines, VSWR, *insertion loss*, and *reflection coefficient*. You may need to refer back to your introductory texts to help with the basics.

When a signal travels any distance along a wire it takes a significant amount of time. The speed of electromagnetic radiation in free space is about 3.3 ns/m. When electromagnetic waves travel along insulated wires the speed drops to more like 5 ns/m, the exact speed being primarily determined by the dielectric constant of the insulator.

If the time taken to travel along the wire {*propagation delay*; *transit time*} is less than 1/20th the risetime of the applied edge, the wire is considered *short*; this idea is emphasised by saying that the wire is *electrically short*. If the transit time is greater than half the risetime of the applied edge, the wire is *long*.

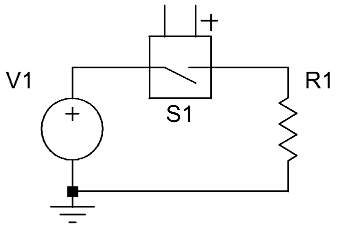
***EX 16.4.2:** A PCB track from an ECL gate is 3 cm long. The ECL device has a 200 ps risetime.

a) Estimate the propagation delay down the track.
b) Is this a *short* track?

High speed digital circuits are really analog. You should *expect* to have to deal with edge speeds faster than 1 ns if you are working with state-of-the-art electronics. The latest SiGe {silicon germanium} ECL devices[†] run at clock frequencies up to 10 GHz with edge speeds of 50 ps.

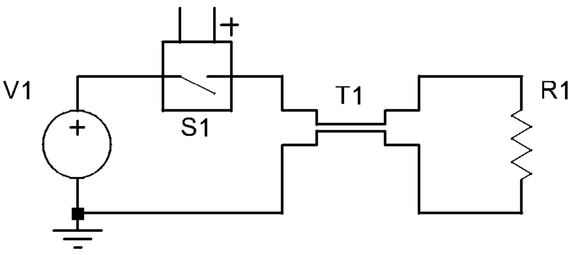
The conclusion is that even a physically short track can behave as a transmission line if the signals are fast enough. Notice that one wire cannot behave as a transmission line; there have to be two wires. The *return* wire may not be apparent, and in this case you will immediately know that there is a problem with the signal routing paths. A high-speed electrical circuit needs *simultaneous* go and return paths.

FIGURE 16.4A:



Elementary courses encourage you to think like this: When switch S1 closes the current flows through the switch, around to R1, through R1 and back to ground. That is fine for DC. With fast edges, or for operation above say 10 MHz, that is not a good model of the real behaviour. You actually have to *think* differently.

FIGURE 16.4B:



When the switch closes, a wavefront starts to travel down the transmission line T1. Current is flowing into the top conductor in the transmission line. No current has yet come out of the other end as energy cannot travel faster than the speed of light. Nevertheless, the same current has to simultaneously flow *out* of the grounded lower conductor on the left side of the circuit diagram. This instantaneous flow of current is due to the *surge impedance* of the line, now more commonly known as the *characteristic impedance*.

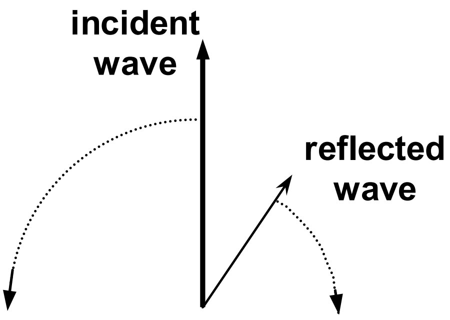
The voltage generator does not *know* what is at the end of the line until the reflected signal arrives. This *tells it* what is at the other end. Until the reflection arrives, the generator only sees the surge impedance of the line.

Now consider steady-state AC. There is still a forward travelling wave and a reverse travelling wave. When the steady-state condition has been reached there is no consideration of multiple reflections. There is only the steady pattern formed by summation of the forward wave (the *incident wave*) and reverse wave (the *reflected wave*). Since voltages are easier to measure than currents, consider the line in terms of the incident voltage and the reflected voltage. At any point on the line the voltage is the phasor {vector} sum of the incident and reflected voltages.

If there is a reflected wave at all, this is not optimum. You are trying to send power down the cable to a load, and some of it is coming back. With a linear line and load, a sinusoidal incident wave gives a sinusoidal reflected wave. Adding two sinusoidal waveforms, regardless of their amplitude and phase, gives another sinusoid.

[†] ON Semiconductor NBSG53A

FIGURE 16.4C:



Moving away from the termination at the far end of the transmission line, the incident and reflected waves on this phasor diagram rotate in opposite directions. They therefore point in the same direction (add) at specific locations on the line, and they point in opposite directions (subtract) at other specific positions. These positions on the line are the maxima and minima in the standing wave pattern. It is very important to realise that if the incident and reflected waves are in-phase (or in anti-phase) then the line appears resistive at that point.

This standing wave pattern is both real and measurable. A *slotted line* consists of a coaxial air-spaced line with an axial slot cut in the solid outer shell, running along the length of the line. A detector probe is inserted into the slot and senses the inner conductor. The probe can then be slid along the line to measure the amplitude and position of the standing wave pattern.

The ratio of maximum to minimum signal is known as the Voltage Standing Wave Ratio [VSWR]. You may see SWR used in place of VSWR, the terms being identical. The reason for explicitly stating "voltage" is that it is also possible to consider standing wave ratios for current and power. In modern usage these other terms are almost never used.

The reflected wave is always less than or equal to the incident wave (unless there is a source of power at the termination). Now the reflected wave is the incident wave multiplied by the reflection coefficient [by definition]. The standing wave maximum occurs when the reflected signal adds to the incident signal, $V_{\max} = V_{inc} \times (1+|\Gamma|)$, reflection coefficient being represented by the upper case Greek letter gamma, Γ. Likewise for the minimum $V_{\min} = V_{inc} \times (1-|\Gamma|)$. The VSWR can then be expressed as:

$$\boxed{VSWR = \frac{V_{\max}}{V_{\min}} = \frac{1+|\Gamma|}{1-|\Gamma|} = \frac{1+\rho}{1-\rho}}$$

ρ is the reflection coefficient magnitude

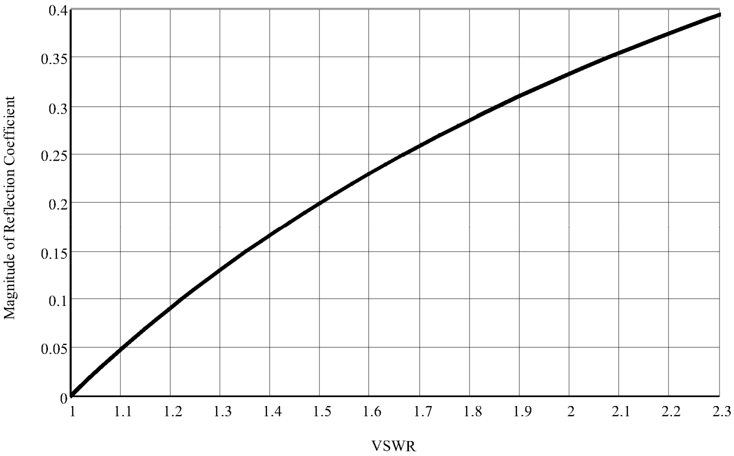
This equation is used to get VSWR from reflection coefficient, even for inappropriate circumstances. When given "output VSWR" for signal sources it is necessary to convert back to reflection coefficient magnitude, ρ.

$$VSWR(1-\rho) = 1+\rho \quad \rightarrow \quad VSWR - 1 = \rho \cdot (VSWR+1)$$

$$\boxed{|\Gamma| \equiv \rho = \frac{VSWR-1}{VSWR+1}}$$

Unless otherwise stated, 'reflection coefficient' means *voltage reflection coefficient.*

FIGURE 16.4D:



The graph shows that for VSWR values below 1.1, the reflection coefficient is approximately:

$$\boxed{\rho = \left|\Gamma\right| \approx \frac{VSWR - 1}{2}}$$

In the same way:

$$\boxed{VSWR \approx 1 + 2 \times \left|\Gamma\right|}$$

Neither VSWR nor $|\Gamma|$ contain any *phase* information, preventing exact calculations from being done; only the worst case can be determined. The worst case is always either a maximum or a minimum, and that is resistive; what looks like a problem in complex arithmetic $\{a + j \cdot b\}$, reduces to a simple resistive divider.

$$\boxed{\Gamma \equiv \frac{Z_L - Z_0}{Z_L + Z_0}}$$

The resistor values used in the worst case analysis are extracted from the definition of reflection coefficient, where Z_L is the load impedance and Z_O is the characteristic impedance of the line.

@EX 16.4.3:

a) If $Z_L > Z_0$ and it is resistive, what is the VSWR in terms of Z_L and Z_0.
b) If $Z_L < Z_0$ and it is resistive, what is the VSWR in terms of Z_L and Z_0.

Return Loss is another way of expressing the magnitude of the reflection coefficient. Certain calculations naturally work more readily in terms of return loss.

$$\boxed{\text{return loss} = -20 \cdot \log_{10}\left(\left|\Gamma\right|\right) = -20 \cdot \log_{10}(\rho)}$$

If a spectrum analyser has a return loss of 9 dB and you put a (perfect) 10 dB ***pad*** {attenuator} at its input, the incident signal is reduced by 10 dB, but the reflected signal is reduced by 20 dB. The return loss from the spectrum analyser input itself is still 9 dB, but it is exposed to 10 dB less of the incident signal. Additionally, the reflected signal is attenuated by the 10 dB pad on the way back to the signal source. The result is that the return loss from the combination of the 10 dB pad and spectrum analyser input is now 29 dB, a considerable improvement. The benefit of the use of return loss should now be clear; since it is expressed in dB, the dB values in the path can be added.

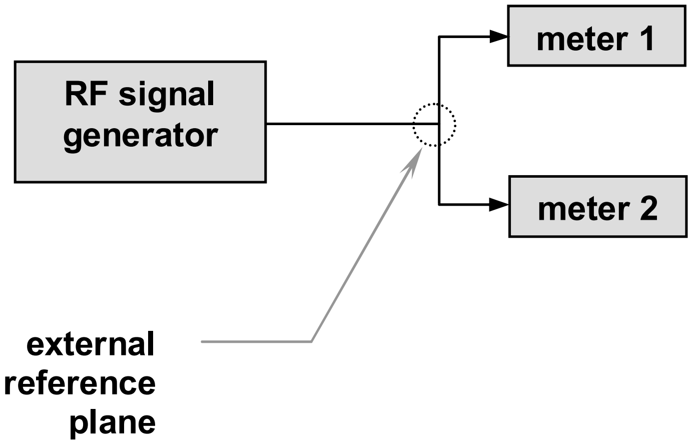
EX 16.4.4: Your scope has a VSWR of 2.1 at the frequency of interest. This is causing you problems. Your boss has given you a (perfect) 6 dB pad to put on the front of the scope to stop you complaining. What is the VSWR as seen at the input to the pad?

EX 16.4.5: An RF power probe has a VSWR of 1.03 at the frequency of interest. Somebody has put an attenuator on the probe to further improve its VSWR. By measuring the power meter's calibrator signal with & without the attenuator you have established that it is a 3 dB attenuator. What is the resultant VSWR of the combination?

In order to manufacture measuring devices for RF work you need to be able to calibrate them. This is major topic on its own. However, there are two basic ways that a measuring device can be calibrated; one could be called *incident voltage calibration* and the other could be called *external reference plane calibration*.

In external reference plane calibration, the idea is to connect two measuring devices by a T-piece {a lossless cable splitting device with three connectors} and 'short' cables. The scheme relies on the shortness of the interconnecting cables to assert that the voltage impressed on one measuring device is the same as that on the other measuring device [let's just call it a meter for simplicity]. By analogy with a lower frequency system, one might say that the terminal voltage has been fixed and the two meters have been wired in parallel.

FIGURE 16.4E:



By 'reference plane' is meant a perpendicular plane through the coaxial cable. In the diagram, the single line connecting the signal generator to the meters is a coaxial cable. The reason for the term 'plane' is that a coaxial connector has both *inner* and *outer* conductors. In practice a *T-piece* would be used to split the coax cable two ways.

On ordinary RF connectors such as type-*N*, BNC and SMA, the inner and outer conductors do not end at the same point. Hence a special fixture is needed to generate a correct 'open-circuit' condition. Imagine an open-circuit connection as being what you would get if you sliced cleanly through a coaxial cable; then go and pick up a BNC, type-*N* or SMA and physically look at the pins and outer housing. They don't look like a cable sawn in half do they?

Using the external reference plane method, if one of the meters is calibrated then the other can be calibrated against it, and the RF signal generator is not a contributing factor. This is a method that can be used to calibrate RF power meters (but not microwave power meters). It is a very easy method to use and since RF power meters have excellent VSWR characteristics, it does not cause excessive measurement errors when the equipment is put into service.

In DC terms you would say that the two meters are connected in parallel and that this connection scheme would give no error. There are errors involved in doing the calibration this way, however, and they depend on the quality of the T-piece, the length of the 'short' cables, the characteristic impedance of the short cables and the VSWR of the meters. Hopefully both meters will have the same type of connectors and this will be the same type of connector as used on the T-piece. If not then type-to-type adapters are needed and these introduce more uncertainties.

Firstly there is an error in calibrating the meters and secondly there is an error when making a measurement with the meters. This is an unusual concept for engineers used to

DC voltage calibrations. Ordinarily you calibrate a meter, get an uncertainty, and that is all there is to it. For use on 50 Ω systems, however, the uncertainty is different for every different situation, and the uncertainties involved are not small. At DC, the loading effect of the DVM on the source is a factor to consider, but when the source resistance is <100 Ω and the DVM input resistance is >1 GΩ, it is easy to neglect the <0.1 ppm error.

On a 50 Ω system, suppose that the interconnecting cables between the meters are so short that there is no significant difference between the voltages that appear on the respective input terminals. In this case you can adjust the calibration factors of the uncalibrated meter to exactly match the calibrated meter. There is now no difference between the two meters; or is there? [Neglecting measurement noise.]

The difference is the input VSWR of the meters. If VSWR is still confusing you, think in terms of input impedance; the input impedance of the meters is different. Ideally the meter would have an input resistance of 50 Ω at all frequencies. In practice the input impedance changes with frequency. This impedance change is characterised in terms of the VSWR changing with frequency. Since VSWR only contains magnitude information, the phase angle of the impedance is unknown, so you can't make an equivalent circuit. What you can say is that if the input VSWR is say 1.1, then the impedance could be 45.5 Ω resistive, 55 Ω resistive or some value of resistance and reactance that also gives a VSWR of 1.1.

To establish what impedances give any particular value of VSWR, draw a circle on a Smith chart, with its centre at the centre of the Smith chart $\left[\frac{R}{Z_0}=1;\ \frac{jX}{Z_0}=0\right]$. The circle will cut the resistive axis at points VSWR and 1/VSWR. Note that impedance values on a Smith chart are normalised so that 20 Ω in a 50 Ω system is plotted as 20/50 = 0.4.

***EX 16.4.6**: Suppose one of the meters has a VSWR of 1 and the other has a VSWR of 1.05. They have been calibrated against each other (external reference plane calibration) and adjusted to give identical calibration factors. If they are both used to measure a perfect 50 Ω generator (one at a time) what is the maximum difference in readings that could be obtained if the meters are measuring voltage.

***EX 16.4.7**: Rework that problem with the meters measuring power.

The problem with external reference plane calibration is that you get an additional error term to evaluate when using the equipment. For this reason RF voltage measuring equipment [50 Ω / 75 Ω] is best calibrated by the *incident voltage* method.

In the incident voltage method the voltage on the input terminals of the measuring device is not directly considered. Ideally the calibration would be done from a source having a perfect 50 Ω resistance and the VSWR of the meter is not important at that stage. The meter's input VSWR is effectively included in the calibration constants for the meter. This makes the meter easier to use because there is no additional error term when measuring a perfect Z_0-matched source.

EX 16.4.8: A voltage measuring meter has been calibrated by the incident voltage method. It has a VSWR 1.5 and is used to measure a source with a VSWR of 2.2. What additional measurement uncertainty should be added to the calibration uncertainty?

If the AC voltage amplitude is supposed to be constant with frequency then the deviation from constant amplitude is referred to as a *flatness error*. [If you are looking on a scope at a rectangular waveform then any bumps and wiggles in the top or bottom of the waveform can also be referred to as ‘flatness errors’.]

Looking into a long lossless cable attached to a mis-matched resistive load, the VSWR of the load is seen as amplitude variation (in the frequency domain) at the input to the cable. The ratio of maximum (RMS) amplitude to minimum (RMS) amplitude is approximately equal to the VSWR of the load, provided the driving source is fairly well matched to the line. However these amplitude variations are not ordinarily measured. It is the amplitude variations at the load itself which are more important. These amplitude variations are a function of both the load VSWR and the source VSWR, although it is more convenient to work in terms of source and load reflection coefficients. The per-unit amplitude variations are a maximum of $\frac{1}{1-\rho_S\rho_L}$ and a minimum of $\frac{1}{1+\rho_S\rho_L}$. The frequency difference between a maximum amplitude and minimum amplitude point is given by $\Delta f = \frac{1}{4\cdot T_{PD}}$, where T_{PD} is the propagation delay down the cable. In order to guarantee seeing both the maximum and minimum amplitudes, the frequency span needs to be just under double this value: $\Delta f \le \frac{1}{2\cdot T_{PD}}$

This formula is derived as follows: Suppose the incident and reflected waves are in phase at the load, the maximum signal condition. If their relative phase changes by half a cycle they will then be in anti-phase and the amplitude will be a minimum. The phase shift, in cycles, is the product of the frequency and the propagation time, but the incident and reflected waves are travelling in opposite directions so their relative phase shift is doubled for a given propagation time.

Some numbers help to demonstrate the effect. Max to min excursions occur every 2.5 GHz for a 100 ps cable; every 250 MHz for a 1 ns cable; every 50 MHz for a 5 ns cable. A 5 ns cable is approximately 1 m long, depending on the dielectric. Since propagation delay is simply the length, *L*, divided by the velocity, and the velocity is reduced by the dielectric constant, the frequency difference between peaks is $\boxed{\Delta f = \frac{c}{2\cdot L\sqrt{\varepsilon_r}}}$ The equivalent equation for waveguide being $\boxed{\Delta f \approx \frac{c}{2\times L}\times\sqrt{1-\left(\frac{f_C}{f}\right)^2}}$

Measuring Bandwidth

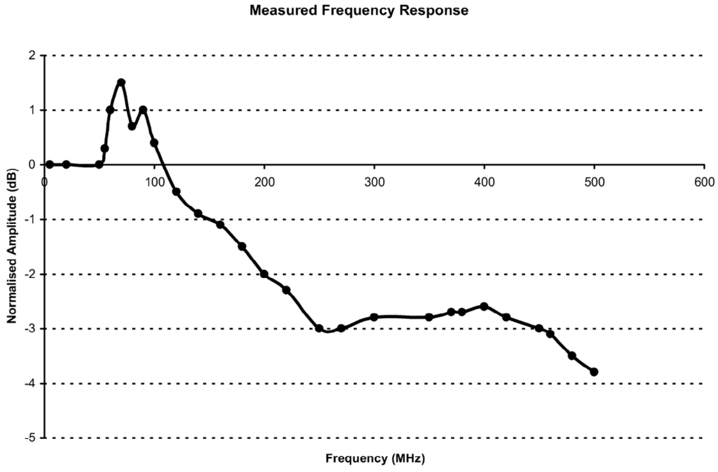
In elementary texts, measuring bandwidth is simple. Feed a sinusoidal signal into the unit and see when the amplitude drops to 3 dB below its LF value. This is an easy task at 100 kHz. You can buy AC calibrators with better than 0.1% flatness and the interconnections will not cause much difficulty. The problems start occurring when you want to measure bandwidths above 10 MHz.

If the unit that you are measuring has a proper RF connector on it, such as a BNC, type-N, SMA &c, then you at least get a defined and repeatable interconnection scheme. If it is a board-level component, you now have to define a jig to mount the component in, then connect up both a test source and a measuring device. Each of these activities adds its own uncertainties, making the final result very uncertain. It may be repeatable, but that is not the same thing as being accurate. Your jig, source and measuring device can introduce sufficient uncertainties on their own to invalidate any measurement you might wish to make.

Perhaps you might think that I have been a bit extreme when I said that that the jig, source and measuring device could invalidate the measurement. Let me justify that statement. When you think of bandwidth, you might be thinking of a nice “well behaved” system like a single-pole RC filter. The roll-off is ***monotonic*** and predictable. There is no guarantee that any real world system or device that you measure will have this sort of response; indeed the whole point of measuring something is to find out what it actually does!

It is not unusual for complicated systems to have complicated frequency domain responses. By this I mean that there are lumps, bumps, wiggles and shelves in the frequency domain transfer function.

FIGURE 16.4F:



This (invented) response has a bandwidth of 250 MHz; or is it 270 MHz; or is it 450 MHz? A very slight measurement error will give a huge error in measured bandwidth on this system. If you only had a quoted bandwidth number for this system, you would not have a good idea of its actual response.

For a system like this one the bandwidth figure is almost worse than useless. It makes you think that you have a quantitative measure of the system when you don’t.

In general you may be interested in the frequency response of filters, amplifiers, voltmeters, spectrum analysers, scopes or other pieces of equipment. For simplicity I am going to assume that they have a smooth monotonic roll-off around the 3 dB point, making a measurement of the bandwidth both sensible and useful.

There are two types of devices to consider. Firstly there is a measurement device such as a voltmeter. In this case it is only necessary to consider the signal going in. Secondly there is a pass-thru device such as a filter. In this case you have to consider not only the input end, but also the output end. I am going to consider the measurement device first as there is only one interconnect to consider. The pass-thru device obviously has twice as many factors to consider, but once the technique has been established, the rest is just number crunching {calculation}.

Before getting too involved in measuring voltages, it is important to know if there is any ‘gearing’ involved in measuring the bandwidth by measuring the voltages. Does a 1% uncertainty in voltage measurement lead to a 1% uncertainty in the measured bandwidth? Consider the transfer function magnitude for an ideal single-pole system:

$$T = \frac{1}{\left|1 + j \cdot \frac{f}{B}\right|} = \left[1 + \left(\frac{f}{B}\right)^2\right]^{-\frac{1}{2}}$$

Differentiate, to get the slope of magnitude response.

$$\frac{dT}{df}=-\frac{1}{2}\left[1+\left(\frac{f}{B}\right)^2\right]^{-\frac{3}{2}}\times\frac{2f}{B^2}$$ The *sensitivity* at the bandwidth point is:

$$\left.\frac{dT/T}{df/f}\right|_{f=B}=\frac{dT}{df}\times\frac{f}{T}=-\frac{f}{B^2}\times\left[1+\left(\frac{f}{B}\right)^2\right]^{-\frac{3}{2}}\times f\left[1+\left(\frac{f}{B}\right)^2\right]^{\frac{1}{2}}$$

$$\therefore\left.\frac{dT/T}{df/f}\right|_{f=B}=-\frac{f^2}{B^2}\times\frac{1}{1+\left(f/B\right)^2}=-\frac{1}{2}\qquad\boxed{\left.\frac{df}{f}\right|_{f=B}=-2\cdot\left.\frac{dT}{T}\right|_{f=B}}$$

If the voltage is in error by 1%, the bandwidth is in error by 2%. This sensitivity is related to the slope of the roll-off. If the slope is steeper, due to more poles, the error will be less. If the slope is shallower, as shown in the previous graph of a system with a 'shelf' in its response, the error will be larger. You could assume that the bandwidth error is at least twice the voltage ratio error in order to avoid having to characterise the response of the system. Having said that, it is sensible to measure the response of the system either side of the 3 dB point to reduce the uncertainty. If your measurement uncertainty is ±1 dB then measure the system response at the −2 dB point. This guarantees that the bandwidth is greater than or equal to the frequency measured. If you need an uncertainty on the measured bandwidth (still with the ±1 dB measurement uncertainty previously measured) then also measure the response to the –4 dB point. The –2 dB and –4 dB points then give the actual bandwidth uncertainty.

I have not mentioned the measurement accuracy of the frequency when measuring the amplitude response, since frequency measurement is not difficult. For frequencies below a few GHz, an uncertainty of better than ±100 ppm is easy, whereas measuring AC voltage amplitude above 10 MHz to an uncertainty of ±1% is somewhere between difficult and impossible. Measuring frequency accurately below 20 GHz is also easy if you have an expensive microwave frequency counter. Again direct reading spectrum analysers can be bought up to 60 GHz, and frequency extension heads can take the spectrum analyser readings beyond 300 GHz. It is all a question of money.

The measurement uncertainty of the frequency would ordinarily be at least two orders of magnitude better than the measurement uncertainty of the amplitude. Hence the frequency uncertainty ordinarily gives a negligible contribution.

16.5 Measuring Frequency

Frequency is a very easy quantity to measure accurately. Inexpensive frequency standards with accuracies better than ±0.1 ppm have been available for decades and always the trend is towards better accuracies. All such standards for general use are based on quartz crystal oscillators.

After having chosen an excellent stability crystal, it is then encased in its own temperature controlled oven. The temperature related drift over a 20°C ambient temperature change is easily reduced below the ±0.005 ppm level. This accuracy easily surpasses those achievable in the measurement of 'electrical quantities' such as voltage, resistance and current outside of national metrology institutes.

To measure the accuracy of a frequency source to better than 1ppm, you might use a 7 digit frequency counter with a suitable spec. The question then arises as to how to check frequency sources and counters amongst themselves.

In the UK there is a *long wave* transmitter at Droitwich transmitting at 198 kHz which has an ultra-stable carrier frequency (accurate in the long term to 2 parts in 10^{11}). This is used as a frequency standard, available to anyone in the country free of charge. Devices which tune in to such a standard can easily achieve specs of ±0.003 ppm accuracy (and better) and are known as *off-air standards*. UK users could also tune in to MSF at 60 kHz. In the USA the WWVB transmitter is also at 60 kHz. Although these low frequency transmissions are fairly insensitive to atmospheric and propagation problems, at the ultra high accuracies required it is still inadvisable to use the system at sunrise or sunset.

Without a 9 digit frequency counter as a ***transfer standard***, it might seem impossible to set up an oscillator to the same accuracy as the off-air standard. The answer is to measure the *difference frequency* between the sources. The small difference frequency can then be measured with relatively low accuracy.

EX 16.5.1:** A particular off-air standard produces a 10 MHz sinusoidal frequency. You are fine-tuning the frequency of your own sinusoidal output 10.0000 MHz crystal oscillator. You have compared the two frequencies using a 6 digit frequency counter and they agree exactly. To get the final trim, you feed the off-air standard and your oscillator into the X- and Y- inputs of a real-time scope set to X-Y mode. You view the resulting ***Lissajous figure and adjust your oscillator to give the minimum speed of rotation of the Lissajous figure.

When you have finished, the Lissajous figure is not rotating at a fixed rate. It is wandering back and forth a bit due to the ***phase noise*** of the two oscillators. All you can say is that the rate of rotation is definitely not faster than one rotation in 25 seconds. What calibration uncertainty can you assign to this arrangement? [Just the uncertainty of the calibration, not the spec of your oscillator.]

In the example, the scope is displaying the difference frequency. Unfortunately this method only works when the Lissajous figure is rotating up to a few cycles per second. Otherwise it is difficult to tell if you are speeding the Lissajous figure up or slowing it down.

Suppose you measure the oscillator with a frequency counter. You can leave them connected together for a while and evaluate the short term stability of the measurement. How much does the reading wander about? This gives a limit on the noise of the measurement. This could be a much better figure than the absolute accuracy of the counter by a factor of perhaps 10×, or even more. If you now do something to the measured equipment such as change its temperature or power supply voltage, you can measure the effect of that change to a much greater accuracy than the accuracy spec of the test equipment.

You can always get better resolution/accuracy on a frequency measurement by comparing two oscillators and measuring the difference frequency, and how it changes with changes to just one of the oscillators. All you have to do is figure out how to measure the difference frequency.

The previous exercise showed the use of a scope to measure the difference frequency such that the scope calibration was not relevant. If you use a digital storage oscilloscope [DSO] the timebase accuracy will be crystal controlled to better than ±100ppm. This is a convenient reference frequency to test your oscillator against.

Provided the DSO has at least as much bandwidth as the oscillator frequency, and provided that the bandwidth doesn't change with timebase, you can try a useful trick. By adjusting the timebase and store size you may get the sampling rate to nominally equal the oscillator frequency, or some sub-multiple thereof. You don't have the facility to measure all frequencies, just a few particular frequencies such as 10 MHz, 1 MHz &c and a few in-between.

If the oscillator frequency is very nearly equal to the scope sampling rate, the difference frequency will be displayed on the screen, regardless of the timebase setting on the scope. Be sure to turn off the glitch detection scheme {also known as peak detection and max-min}; the difference frequency is actually an ***alias***.

EX 16.5.2: Your DSO has a vertical accuracy of 2%, a timebase accuracy of 10 ppm and a bandwidth of 100 MHz. It is set to sample at 5 MS/s. The store length is 10,000 points and the timebase is set to 200 µs/div. The ambient temperature is 23°C ±3°C, so the scope spec is valid. There is a fairly ragged looking sinusoid displayed when you apply an oscillator that is known to be producing approximately 50.0 MHz. The displayed waveform has an amplitude of about 5 div and the period is just over 4 div. Give limits for the frequency of the oscillator.

If you have a frequency to measure that is not at a multiple of a DSO timebase, or if the beat frequency against another oscillator is going to be faster than a few Hertz, another difference frequency measurement method is needed.

You can *down-convert* a signal by just feeding it into a diode, along with a local oscillator signal. However, the sensitivity of this method is not very good and the output has to be heavily filtered to remove the original input signals. A proper commercially made *double balanced mixer* [a passive device consisting of diodes and RF transformers] gives a much larger output signal and also minimises the feedthrough of both the RF and local oscillator signals.

Above a few gigahertz the local oscillator may become prohibitively expensive (Gunn oscillators are available to above 100 GHz but may cost >$3000 each). Possible solutions involve analog *frequency multipliers*. For example a sinusoidal signal put into a non-linear device will produce a double frequency signal, amongst others. The technique would be to inject the fundamental and tune the output circuit to reject everything but the double frequency signal.

By feeding the oscillator under test and the reference oscillator into a double-balanced mixer and low-pass filtering the output, the difference frequency can be observed. This can then be measured on a scope, spectrum analyser, or fed into a sensitive frequency counter. If the difference frequency is below 10 Hz then either a chart recorder or a DSO in *roll mode* can be used. In fact when comparing an off-air standard and a temperature controlled reference oscillator, I have used a DSO in roll mode at 200 seconds per division. You can then see the average trend of the relative drift with a difference period around the 200 second level. Note that $1/200^{th}$ seconds for a 10 MHz oscillator pair is a difference frequency of 5 parts in 10^{10} .

Direct measuring digital frequency meters are available up to 20 GHz. Above this measurements have typically been made with manually operated cavity dip meters. The dip meter is placed in series with a power meter and the dial moved until a dip (around 1 dB) is seen on the power meter. This is a slow process and the accuracy is limited (≈0.1%). High orders multiplier heads can be used with spectrum analysers up to and beyond 300 GHz for faster more accurate measurements.

16.6 Measuring in the Time Domain

Time ***domain*** characterisation of systems is almost universally done by applying a rectangular waveform and seeing what happens to it. The result is usually referred to as the *pulse response* of the system, although the term *step response* is also used.

Learned texts like to consider the *impulse response* of a system. An impulse is a narrow rectangular pulse whose amplitude tends to infinity and whose width tends to zero. A *unit impulse* is defined as one whose product of amplitude and duration is 1 V·s. Thus 1 V for 1 s would be 1 V·s. Likewise 1 MV for 1 μs would also be 1 V·s. Impulse response testing of time domain systems is seldom done, due to practical difficulties.

Consider doing an impulse response test on a scope. In order to be meaningful, the impulse would need to be narrower than the rise-time of the scope. Its amplitude would also need to be well over full scale. The result would then be entirely undefined as the scope would be seriously overloaded. Whilst it is usually acceptable to overload a scope by perhaps one whole screen (8 divisions) at a few kilohertz, it is most certainly not acceptable to overload the front end by this amount for an edge comparable with the risetime of the scope. In practice, many manufacturers do not give a specified response for just a few divisions of overload, even at low frequencies.

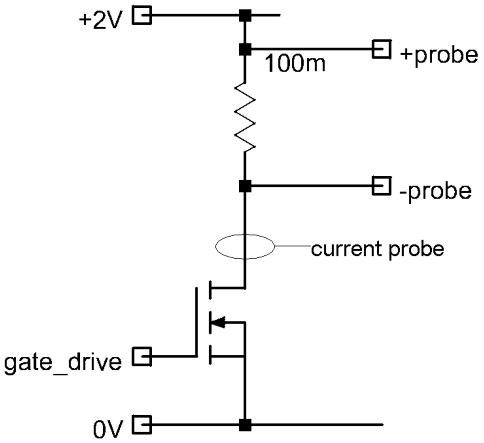
Thus any impulse testing of a device must first be limited by the linearity and overload capabilities of the device. It is for this reason that pulse response testing is the key method used to characterise systems in the time domain.

The response of a system to a pulse is limited to three broad classes:

1) The response overshoots and possibly rings as well.
2) The system gives a simple smooth monotonic response.
3) The system gives a slow settling, possibly ***non-monotonic*** response, with evident multiple time constants.

Depending on the application, any one of these responses may be acceptable, although the first two responses are more usual, and more acceptable. In a system composed of several stages, for example, it is often acceptable to allow one or more stages to individually have some ringing and overshoot, provided the bandwidth of these stages is more than 5× greater than the overall system bandwidth.

FIGURE 16.6A:



This circuit arrangement shows how *not* to test the step response of a current probe. The idea is that a step current is generated by the N-channel MOSFET when it is turned on a by a 12 V 10 Hz square wave from a signal generator. The current is measured directly using a current probe and indirectly by measuring the volt-drop across the 100 mΩ resistor.

The technique given above will easily produce a 10 A current swing in < 2 μs.

***EX 16.6.1**: What is wrong with the above measurement technique for evaluating the step response of a current probe at 10 A, and how can it be improved?

Some people spend lots of money buying expensive equipment to make measurements when all they really needed to do was to first employ some "brainpower" to reduce the measurement difficulty.

Suppose you want to measure a 100 mV signal which is "sitting on" a DC bias of 50 V. The "brute force" approach would be to buy a scope with 16-bit resolution. Since the 100 mV signal is only 0.2% of the overall input signal, the effective measurement resolution is reduced from 16 bit, 1 part in 65536, to less than 1 part in 131. The effective measurement resolution is then more like 7 bit and a great deal of money has been wasted.

An alternative approach is to use a scope with a built-in offset feature. The difficulty with this approach is that if the DC level is not very stable, it may not be possible to back-off the DC level using the scope's offset control.

A better answer is to use a differential probe. This way the DC bias can vary or be noisy, and the CMRR of the probe will tend to reduce this drift or noise of the bias level. If, instead of a differential probe, you use two ordinary probes on two separate scope channels, using the A–B mode on the channels, or indeed A+B with B inverted, you will get a really lousy answer. The difficulty is that each channel has to handle the full signal plus the full common-mode noise. This "differential measurement" has used up all the dynamic range of the system and the result is inadequate signal resolution.

If you need to make such a measurement, and no differential probe or scope with offset is available, do not despair. You can use an ordinary 10:1 passive probe and a BNC input/output circuit box to make up you own DC offset removal circuit. Make sure you always have a BNC circuit box on hand in the lab; they are very handy for building up the little filters, and so forth, that you occasionally need.

A 10:1 scope probe is a 9 MΩ resistor which feeds into the 1 MΩ input of a scope. If you shunt the scope's 1 MΩ by another 1 MΩ you will have made a 20:1 probe. If instead of wiring the 1 MΩ resistor to ground, you connect the resistor to a power supply or DC calibrator, you have made a probe which can give more offset than any scope will ever give you.

Internally link the input to output of the BNC circuit box, plugging the circuit box onto the scope BNC, and connecting the probe to the BNC circuit box. Solder a 1 MΩ resistor to the BNC box link-wire and take the other end of this 1 MΩ resistor, and a ground connection, off to a power supply or DC calibrator. The circuit box will also require a capacitor to ground at the scope input, in order give the probe compensation enough range.

Measuring Clock Jitter

Clock jitter is a measure of the stability of a clock from one cycle to the next, or from one cycle to some period many cycles later. It is a key factor for measurement accuracy in a sampling system and for signal integrity in a communications system. The measurement of clock jitter is therefore important. At frequencies >300 MHz the clock may be sinusoidal; you might therefore look at the jitter in the frequency domain. In this case you would be measuring ***phase noise***.

When looking at a clock in the frequency domain, one should neglect the harmonics. After all, the clock might ideally be square, in which case a generous amount of harmonic content is called for. What you must seek out and destroy are the ***sub-harmonics***. These may be at integer sub-multiples of the clock frequency and therefore cause a repeating pattern of jitter. Since spectrum analysers cannot generate sub-harmonics any seen on a spectrum analyser are really there.

Typically an oscillator output might drive a divider chain, and be used at the fundamental as well. The divider chain can load the oscillator at a sub-harmonic rate and therefore generate sub-harmonic content on the clock. I am not talking about huge effects here. The sub-harmonic signal might be down at –40 dBc, but this can still be large enough to cause problems. The other way these sub-harmonics can be generated is by putting harmonically related signals through gates in the same package. There will inevitably be some degree of cross-coupling between the gates in the IC package, and again the result is sub-harmonic modulation.

Jitter with a fast repetition rate can be seen on a real time scope. However, it is more common to use a DSO to capture all of the jitter positions of the clock. Modern DSOs have various proprietary {own brand} *persistence modes* and such a mode is useful for viewing excessive amounts of jitter. All persistence modes will show the peak-to-peak jitter, the critical figure for digital clock timing data.

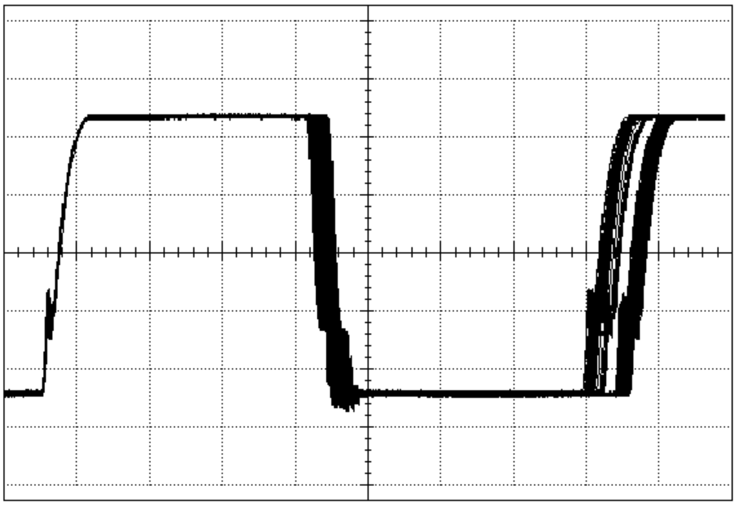
The most obvious way of observing clock jitter is to display the clock on a scope and show two rising edges on the screen.(This whole section could obviously be applied to falling edges instead.) The first rising edge would be the trigger edge and the second would be the start of the next cycle of the clock. By using some sort of persistence mode on the scope the jitter can be readily measured. The first rising edge should not be jittering because it is the triggered edge. Thus any width in the persisted image of the first edge will only be due to trigger jitter on the scope.

Jitter on the second rising edge will include the trigger jitter, but will be larger due to the jitter on the clock. This is a relatively insensitive method of measuring jitter. It would only be suitable for measuring jitter which was greater than about 1% of the clock period. The amount "1% of clock period" is more commonly referred to as "0.01 UI", or 0.01 of Unit Interval.

For this type of measurement the trigger jitter is critically dependent on the noise on the waveform, the quality of the scope trigger circuit and the slope of the rising edges of the clock. If the trigger edge is moving about on the display this is either due to a poor quality scope or excessive noise on the waveform causing mis-triggering.

FIGURE 16.6B:

This is a real persistence plot of the gate drive on a switched-mode power supply. The period of the switcher is not constant and you can see the jitter on the low period is greater than the jitter on the high period. Note that the trigger edge is relatively jitter free.



To get better resolution, trigger delay can be used. The first step is to trigger on the rising edge, using perhaps 10% pre-trigger, and zoom the timebase to see the edge in detail. The triggered edge is now being viewed with a great deal more resolution than was done before. This is a check on the trigger system to see how stable a trigger you can get. The width of the resulting persisted signal can be measured. This gives a measure of the uncertainty in the final result.

Now, trigger delay equal to just under one clock cycle is applied and the jitter on the next clock edge is viewed. If the clock edge is very fast compared to the clock period, let's say that the risetime is less than one hundredth of the clock period, then jitter can be measured down to more like one tenth of the risetime. This means jitter resolution of better than 0.1% UI.

This trigger delay method measures jitter well below 0.1% of the clock period, but only at modest frequencies. The DSO has a finite maximum sampling rate, limiting the jitter that can be measured. Obviously you can't measure jitter that is less than one sample interval of the DSO; or can you? This is all a question of how clever you are, or how clever the software can be made to process the digitised data.

Let's suppose that I have a DSO with a maximum single-shot sampling speed of 1GS/s. Now that is fast, but not the fastest possible speed. The thing is that you might not be able to afford a faster one. The acquired points are 1 ns apart. How can you measure the jitter on a 100 MHz clock [10 ns period] to any degree of accuracy? After all, the 100 MHz clock is probably ultimately referenced to a SAW or crystal, both of which will have an excellent Q and therefore very low jitter. The jitter should certainly be better than 100 ps ptp [which is a huge 1% jitter].

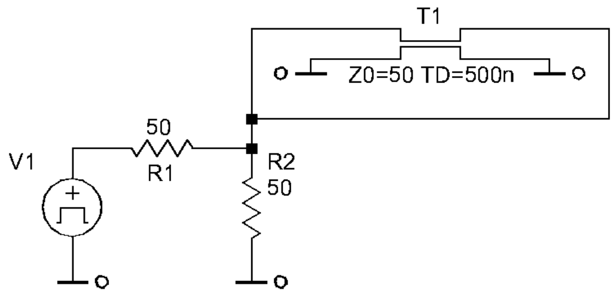
The previous techniques described suffered from trigger jitter. Noise on the slope of the triggered edge causes jitter which increases the uncertainty on the clock jitter. It would be good if this factor could be eliminated. This next technique can only be used when the clock being measured is an integer sub-multiple of the master timebase clock of the DSO. You need to deliberately ***alias*** the external clock oscillator, displaying a difference frequency. Ideally you would set it up so that you sample at one position on the clock waveform then delay by an integer number of clock cycles plus say one 500th of a cycle. If each sample point is taken in this manner you get a nice alias of the clock waveform built up in *one sweep* of the timebase. Thus the trigger jitter is eliminated *completely*.

It *is* possible for the trigger waveform to affect the sampling clock in the scope, however. Since the trigger is not being used in this measurement, it is best to turn the trigger pickoff to an unused channel or move the level away from the trace and just use the AUTO trigger function to invent a trigger every now and then. This minimises the trigger to sample cross-talk.

Another measurement method is to get the DSO automated measurements to measure the waveform's period. Whilst the sampled resolution of the waveform would not give an accurate measure of the waveform period, linear interpolation of the mid-amplitude crossing points does give an accurate measure, particularly when the waveform has a rectangular shape. Using this method, the trigger jitter does not cause a measurement uncertainty. Furthermore, post-processing the waveform period to give the standard deviation of the measurement yields the RMS period jitter, the two terms being almost identical (see the appendix).

There are of course specialist pieces of test equipment specifically designed to measure jitter and/or phase noise if your budget can stretch that far. In any case, you will want to qualify your test method against a jitter-free clock.

FIGURE 16.6C:



V1 and R1 are the clock source. R2 is the jitter test equipment input. T1 is a coaxial ring delay line, which could be made from a piece of coaxial cable or from semi-rigid coax. The key point about this system is that the pulse generator has to create a rectangular pulse narrower than the length of the delay line. In this case a single input pulse creates two rectangular pulses, spaced apart by the length of the delay line (500 ns in the diagram above). If the generator produces 1 V pulses without the ring delay line, both pulses are 0.5 V when using the ring. Measure the jitter between the two rising edges (or the two falling edges) and you have a jitter-free source. (See also page 388.)

16.7 Measuring Interference (noise)

There are two specific types of interference that are encountered in "noise debug" situations; *synchronous* and *pseudorandom*. Just what you do after you have measured a noise source is another matter, and that is left to a later chapter. For now I want to just look at the process of making the measurement.

By *synchronous* noise I mean that the interfering source is locked in frequency or time with the resultant noise signal. The time difference between the two will not be zero, but it should be relatively fixed. Synchronous noise is generated by an interfering source and couples to the *victim* circuit in a variety of ways. These ways include:

☹ capacitive coupling
☹ magnetic coupling
☹ electro-magnetic coupling
☹ ground loop coupling, and other common-impedance coupling
☹ supply rejection coupling

Regardless of the coupling mechanism, the noise appears at the output of an amplifier or signal conditioner. The task is now to measure it. The first step is to guess what it might be! This may seem a bit backwards; after all, you are looking at the amplifier output in order to see what noise is there. You may find it strange to guess what noise is there before measuring it.

That is the way these things work. You need to know what it is you are looking for before you try to measure it. If you are looking for 2 GHz clock harmonic noise you will not measure it with a 3½ digit hand-held multimeter. If you are looking for noise related to LEDs being switched on the front panel then you will not find it using an RF spectrum analyser. You must use the right tool for the right job, and with the right connection scheme.

So, you need to know what the noise is before you measure it and yet you have not made a measurement yet. How should you proceed? Well, the thing to do is to make an *educated guess*. If there are any switched-mode supplies in the system then you can guess that there may be noise due to them on the sensitive analogue circuitry. Guess that it exists, measure it, and then you will be able to say if it is too large or not.

I have told you before that no effect is zero. If there is a switched-mode supply anywhere in the system, the effect on the analog signals will not be zero. It is a question of making your measurement technique sensitive enough to be able to detect undesirable levels of this interference.

There are two basic measuring tools to be used in your noise measurement kit; the DSO and the spectrum analyser. Each has it strengths and you do need both. Let us first continue with the case of the switched-mode power supply. Switched-mode power supplies emit electric and magnetic fields, regardless of how well screened they are. This is an excellent source of a trigger for your DSO. If the switcher is in a metal case then just take an ordinary 10:1 probe and poke it through any ventilation hole in the metal case, ideally next to one of the magnetic cores. It is unlikely that you will fail to get a big stable trigger signal.

Variable frequency switched-mode supplies are a nuisance in this respect, but if you change the timebase so that you get only one cycle of the trigger waveform on the screen you will be able to see what is happening on the switching edges of the power supply. Now you can put another channel of the DSO onto the victim amplifier output. The noise due to the switched-mode supply will be locked relative to the trigger signal, although the phase shift relative to the trigger will be arbitrary.

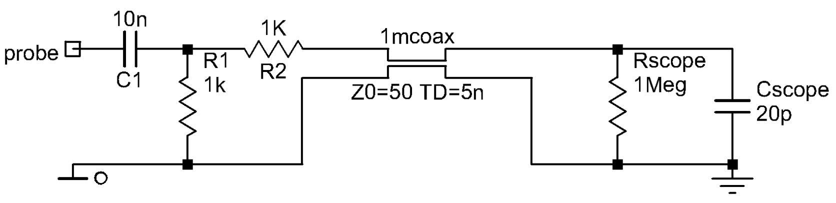
If the noise is not apparent and the total noise appears entirely random then turn on averaging. Turn the averaging up to 64 or higher and the other noise sources will just fade away, leaving only the noise due to the switched-mode supply. (This is why you need a DSO. You can't do averaging like this on an analog scope.) If the noise due to the switcher is very much less than the other noise sources, you will have to turn the averaging up even higher to get a stable measurement. Using this method you will be able to say with certainty that the noise on this output, under these conditions, is less than a certain number. If you write that the noise is zero then you certainly need to restudy this book from the beginning! All you should say is that the interference from the source you are measuring is less than whatever the measurement noise and resolution permit you to state with certainty. It is also possible that you will need to repeat this measurement after other interference sources have been killed off; they may have been masking this interference source, or the process of reducing those other interference sources may have increased the level from this one.

One important point relates to the sensitivity setting of the scope. Adjust the scope's sensitivity so that the noisy signal fills the screen without clipping before averaging is turned on. Having turned on averaging, *do not then increase the sensitivity*. The scope is digitally processing the acquired data to produce the averaged waveform. If the sensitivity is increased so that the data prior to averaging is clipping, the resulting waveform will not be a correctly averaged waveform.

Now I have indicated how to probe a circuit in the section on probes and probing. However, you may need to go further than that. Let's suppose that you are looking for noise around the millivolt level 'standing on' a DC level of 10 V. It would be obvious to anyone that you would select AC coupling on the scope in order to get more sensitivity on the reading. Clearly on a 10 V signal you would otherwise be scaled to 1 V/div with the ground shifted off the screen. With this setting you could not hope to see a 1 mV level. By AC coupling you can use a 1 mV/div or 2 mV/div setting and easily see the noise.

Ok; that is very straightforward. But what if the noise source is 1 mV 50 kHz and it is standing on a 200 mV 120 Hz signal? You have to set the scope to perhaps 50 mV/div and there is again no chance of seeing the noise source you are looking for. You need to enhance you measurement technique by making your own highpass network. This is not hard below 10 MHz. All I am talking about is soldering an RC network to the board under test to make a suitable filter.

FIGURE 16.7A:



With this "probe" you can look for noise from a switcher at 30 kHz. There are only three components to put in place; C1, R1 and R2. These would be wire-ended parts, soldered to the bare end of a coax cable. This gives a very inexpensive, but still accurate probing method which will measure the desired noise signal without passing all the other interfering sources. You will need to use probing techniques such as this if you ever have to get the measured noise level down in the presence of other noise sources.

In order to get the lowest possible noise level in any system, you need to make a definite effort to check each possible noise source and ensure that it is low enough. For a digital acquisition system, less than 0.2 LSB ptp would generally be considered as acceptable for any individual noise source, but you should appreciate that if you have 5 of these uncorrelated sources then there will be a ptp noise of 1 LSB, and this may be unacceptable. (If you have a 16 bit system than it may be that several LSBs of noise are acceptable because the random noise may be many times larger than this.)

Another reason for sub LSB noise requirement is that if you FFT the data, you can see much lower than 1 LSB noise levels. Consider a 12-bit system. 1 LSB is 1 part in 4096 levels, which corresponds to –72 dBFS. But the noise floor on a 4096 point long FFT of 12-bit data should be more like –100 dBFS. In fact the longer the FFT, the lower the noise floor. The quantisation noise and the random noise get spread out to fill up the FFT. This means that you can see sub-LSB signals, *but only in the presence of other signals*. If the signal just happens to sit nicely in the middle of one quantisation level, then the sub-

LSB resolution is no longer available. Applying a small-signal allows the sub-LSB resolution to reappear, and this is strongly related to the subject of ***dithering***.

Noise sources to test for include:

- ☹ Mains frequency (50 Hz or 60 Hz)
- ☹ All switched-mode frequencies used (including display backlights)
- ☹ Fan currents (DC fans can generate unpleasant current waveforms).
- ☹ Power supply repetitive load current patterns (bursts of activity of RAM for example can produce periodic dips in the power rails and hence synchronous noise).
- ☹ Relays, motors, solenoids or other high current loads.
- ☹ DAC/ADC strobe/enable lines

Notice that I am saying that you should use the mains frequency to check for noise. It is true that most noise will be at double this frequency, but if you use the lower frequency you will see noise of that frequency and any higher harmonic. Hence more problem areas will be highlighted if the fundamental frequency is chosen.

The averaging technique mentioned above applies to single-ended 10:1 probes, active probes, differential probes and isolated probes. The extraction of a *correlated signal* from a mush of other asynchronous noise is a truly powerful measurement technique which you must become familiar with and be able to use effectively.

Having seen the idea of coupling mechanisms from the previous paragraphs, you should now be wondering what on earth *pseudo-random interference* is. The name suggests that there is no cause for the interference! The point I am trying to make by the use of this name is that *you cannot find a trigger source* to use for the averaging process.

In systems with microprocessors, the general address/data bus activity is a pseudo-random noise source, and it will inevitably get into the analog circuitry to some extent. The best way of looking at this noise is by using a spectrum analyser or a DSO in FFT mode. In practice what happens is that there is a broad band ‘mush’ on the analog output {apparently random increased noise level seen over a wide range of the spectrum analyser display} and the use of appropriate decoupling and ground modifications will tend to lower the mush over a broad band of frequencies.

The best way of coupling into the spectrum analyser is via a thin coax cable soldered directly to the circuit under test. Since a spectrum analyser has a 50 Ω input resistance, and will be destroyed by as little as 5 V of signal, it is wise to put a 10 nF coupling capacitor in series with the cable to prevent destruction of the input mixer, a *very* costly thing to repair. It is often necessary to also put 100 Ω - 1 kΩ in series with the cable, since your circuit would otherwise have to drive the 50 Ω load of the spectrum analyser input.

Having measured the pseudo-random noise, the key thing to do is to change the equipment/system setup and see if you can increase or decrease the level. If, for example, you can get the system to do more (or less) RAM writes, and this changes the amount of pseudo-random noise, you have immediately established the source of the noise.

CH17: design principles

17.1 Packaged Solutions

This book does not contain a series of electronic sub-circuits for you to incorporate into your designs. This sort of information is already widely [and *freely*] available in application notes from companies like:

✓	Analog Devices	www.analog.com
✓	Avago /Agilent / HP	www.agilent.com
✓	Linear Technology	www.linear.com
✓	Maxim	www.maxim-ic.com
✓	National Semiconductor	www.national.com
✓	NXP / Philips	www.nxp.com
✓	Texas Instruments	www.ti.com

and others too numerous to mention. You must use this applications information as a resource or you will never produce optimum designs. Just treat the application information as a starting point, however. Don't expect their information to be totally correct or totally complete; you will be disappointed. And don't expect the manufacturer to state the weak points of their products. You may have to look at competitor's products in order to find out about the products from the manufacturer you are interested in!

The more packaged solutions you use in your designs, the quicker you will be able to produce products and the happier your bosses will be. This then gives you more time to work on the really difficult design challenges.

I introduced the idea of a design as a commodity in Chapter 2. You have to sell your design, first to your bosses, and then to an end-user {customer}. Why should they buy *your* design? The answer is that your design has to be better than the competition. Let me give you an analogy: In motorcycle races the difference between a supreme champion and a supremely useless rider can be a matter of just pushing too hard. The supreme rider goes around corners fast; as fast as he can. He takes an optimum line {path} through a corner and the tyres retain their grip, but only just. The supremely useless rider always pushes it just that little bit too hard and keeps skidding off the track. The point is that if you want to stay out ahead you have to push hard, but not so hard that you come off {crash}. It can be a fine line to tread!

How does this relate to design? It's all a question of safety margin. If you put in lots of safety margin then your designs will all work first time; that may be what is needed. It really depends on the market you are serving. For high volume commercial and industrial use you will get sacked {laid off; let go} as you will be producing designs which are too large, too expensive, too complicated, too heavy, too … [anything bad]. This is not the only way to stay ahead. You can also do so by having a technological advantage. If you spend development time and money on increasing your intellectual property {know-how} or developing custom chip-sets then you can also gain a significant competitive edge. Realistically you will need to do all these things.

The result can be that *your designs will not work first time*. If you have an expectation of state-of-the-art designs working as drawn then you will be sadly mistaken. Now remember that this is a book on *analog* electronics not digital electronics. With modern simulation tools, I fully expect a semi-custom or full-custom digital IC to work correctly the first time. What I don't expect is that I will wire up a new amplifier and immediately find that the noise, pulse response, bandwidth, ENOB, SFDR, ***THD+N***, or whatever, will automatically be correct. These will need adjustment, tuning and development to get right.

This idea is most especially true when you are doing something new (for you). If this is the first 16-bit ADC buffer you have made, the previous ones being 12-bit, you should expect that you will have missed some points that will make the design less ideal than you might have hoped. Thus time-scales must reflect doing at least two, if not three iterations of the PCB, in order to get a reasonable chance of meeting both the spec and the timescale.

Even apparently simple design tasks often take two iterations of the PCB to get right. It may be better [quicker and cheaper] to allow for two iterations of the PCB to get the design right rather than endlessly simulating and modelling to ensure it is right first time. If your performance is monitored by your boss on the number of "first time successes", recognise that this is a far from ideal measure.

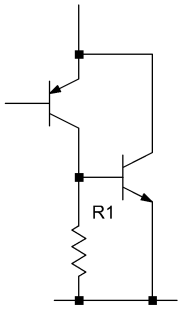
17.2 Custom Analog Chip design

Analog ICs have to work first time or they have to have enough test structures and test points in them to find out what is wrong. The situation is this: the chip comes back and it is not working. What do you do now? Clearly there is something wrong, and if it is something that is buried deep in the middle of the IC then you are unable to determine what the exact fault is. How can you correct it, if you don't know what the problem is?

Now understand that I have been to this dark place and returned successfully, but it was more by luck than good planning. I had a switched gain chip delivered from the fabrication plant and it was oscillating internally at some modest frequency. How did I know that it was internal? Well I shorted the inputs, unloaded the outputs, reduced the gain to a minimum and I could still see the oscillation.

On this semiconductor process the PNP transistors were next to useless. Low current gain (×10), low current capability (1 mA) and low f_t (1 MHz) compared to a current gain of 100, a current rating of 60 mA, and an f_t of 500 MHz for the NPNs. This was an inexpensive process for automotive applications and was not state-of-the-art. The way to get a better PNP was to make a compound device.

FIGURE 17.2A:



This compound PNP makes up for the lack of gain and current handling capability in the basic PNP device. R1 sets the gain of the loop. Put too high a value in there and the circuit will oscillate. Now obviously I tested this with max and min SPICE models, but SPICE will not locate potential oscillations without help (see §14.7). In the main path an AC analysis would have shown severe peaking of the gain. This bias path was not excited by the full signal so its frequency response was not evident.

This sub-circuit set a bias point deep inside the chip. How did I prove it was the culprit?

I just looked at the circuit and that was the only feedback mechanism I could see. To prove it I took the metal lid off the package and probed the die under a microscope. This would not be possible on a newer faster chip. This was mature (old) technology back in 1989 and I just managed to hand probe that area.

The probe shorted out something in that area and the oscillation stopped. Success! I was lucky, the point was proved and the redesign consisted of just reducing the loop-gain by reducing the resistor. Nowadays you would have to get the chip probed with a specialist jig. The moral of the story is to have intermediate points to test in any design. Also use the techniques of §14.7 to test the individual stages.

You could put a few extra test pads on the chip (that don't have passivation over them). You could then find out why your chip is not functioning. This is particularly necessary if you are working with new technology that is not fully characterised.

Alternatively, make a few test blocks that only have part of the complete design in them. This way you can see which stage is not doing its job. Obviously there will be some circuits that cannot stand the extra capacitance caused by test nodes. The test nodes would also increase the cost by using up more die area. Just don't leave yourself in the situation where you plug the chip in and it either works or it doesn't. You have nowhere to go in this position because you have no information to tell you why the previous iteration didn't do what you expected. This should be considered an essential part of any complex design.

17.3 Mature Technology

Technology is moving forward and what you will find is that rather than make a circuit out of a few transistors, a few diodes and some R's and C's, you get a functionally better part for less money, using up less board space, if you pick an encapsulated function from one of the volume semiconductor manufacturers. They are encapsulating more and more functions, de-skilling the whole business of analog design.

Take voltage regulators, for example. If you want a simple voltage regulator then you would not make one out of discrete parts. That would be absurd. You can buy one that is fully specified and short-circuit protected for about the same cost as the high power series pass transistor that you would need to design your own regulator. It is all a question of production volume. The manufacturers are making so many of these regulators that the cost is extremely low.

However, this is for mature {older design} parts. If you are buying a voltage regulator such as a 7812 [12 V positive voltage regulator] which was designed decades ago, there is no money needed by the supplier to cover the design costs, so the part is really cheap. Up until about 2002 the newer ***LDO*** voltage regulators were often five to ten times the price of the older designs.

It is for this reason that every time you come to start a new design, you have to be familiar with not only the technology available from the semiconductor suppliers, but also what their prices are. This is a key decision making point.

Perhaps you feel let down. What is the point of learning all this stuff about resistors, inductors, transistors &c if you are just going to stick-in {use} a packaged solution? The thing is that whilst packaged solutions are a vital part of your design kit, they will never fit into all the nooks and crannies {obscure places} in your design. The packaged solutions are designed for the high volume market, so common features of the design are encapsulated. Features that are highly specific for your application are not necessarily going to be represented. Also, some things are very hard to do in integrated circuit form, particularly when they involve high voltage and/or high power.

Until perhaps 1995 it was usual for the inputs of an IC to be constrained to be not more than about 0.5 V outside of its power rails. This was due to junction isolation and clamp diodes. ICs are now available that can have inputs 100 V above the power rails. [This would be where there are precision internal resistors wired as potential dividers.]

Hence the analog designer will find that routine tasks can be done by using packaged solutions very quickly and cheaply. As soon as the packaged solutions are not available then the design time [and therefore the design cost] increases by orders of magnitude. Yes, order*s* of magnitude. If you want a voltage regulator which can operate with 3 V headroom, needs to supply 120 mA and supplies 12 V ± 10% then you can drop in an off-the-shelf regulator. Design time probably less than ten minutes.

If you want a 110 V ± 0.1% regulator at 1 A then you will be on your own. First you sketch out the circuit with values and types, then prototype it and add a few R's and C's to enhance its performance [stop it oscillating!]. Then you need to characterise its supply rejection, optimise that by changing R's and C's, measure its TC, check its short-circuit capability. The elapsed time could be a few weeks, since you might have to order the parts, get in a queue for somebody to build it, wait for a temperature controlled oven to become available &c. The design time is certainly going to be >100× the previous design. Bosses can become impatient with this because they know that it only took 1 week to design this whole page of circuitry over there and yet you have spent a month working on a little bit of circuitry up one corner of the page. This is a question of education, or lack of it, in the boss. It falls to you to re-educate your boss. Simple looking analog circuits consisting of only a few transistors, resistors and the like can take an apparently unreasonable amount of time to get working.

The other point is the skill involved. The less analog design you do, the longer it will take when you really have to get deeply involved in something complex. If you were designing discrete voltage regulators all the time you would get good at it. However, what is more likely is that this one will be the first one you have ever designed. The ability to learn new techniques from books and application notes is therefore crucial to your future success.

17.4 Gain and Bandwidth

Let's look at the rules of amplification. Signal amplification costs money. More amplification cost more money. Amplification at higher frequencies costs more and is less accurate than amplification at lower frequencies. All a bit qualitative, I know. Let's try a few examples.

If you want a voltage gain of 10× at a frequency up to 10 kHz then this is easy. You can do it with a 1 MHz gain-bandwidth product opamp and a couple of resistors. If you want a gain of 15× then it would cost the same; the opamp could still handle it. However, if you want a gain of 10× at 100 kHz then the opamp is running up to its limit and probably beyond its limit under worst case tolerances. You need to buy a higher gain-bandwidth product opamp and that costs a little more money. More gain or more bandwidth costs more in steps, rather than as a continuous function.

You can keep buying more expensive opamps. They do get progressively more difficult to use, however, because they need better decoupling of the power supplies and better layout to get them to work optimally. As of c.2004 you could get voltage feedback bipolar opamps with gain-bandwidth products up to around 1.5 GHz.[†] They run from ±5 V power rails, rather than ±15 V rails, because higher frequency transistors are smaller and can therefore only withstand small voltages. They are also often only stable with gains of 5 or 10, meaning that closed-loop bandwidths around 300 MHz are the present limits. Current feedback types can run to higher closed loop bandwidths, say 600 MHz, although unity-gain buffers are available up to 1 GHz.

The gain is defined by a pair of resistors at low frequencies. As the frequency increases, it is necessary to add capacitors to the circuit so that the high frequency gain is set by the capacitors (voltage feedback types only).

When opamps can't achieve the required bandwidth you have to change technology. This used to mean making a discrete transistor amplifier, but this option is now effectively closed. Technology in opamps has moved on to the point that a discrete solution would need to be running above 300 MHz. It is really not very sensible trying to make high gain amplifiers at >300 MHz in discrete components. The stray capacitance and inductance is very limiting to the circuit performance. There are two possibilities:

- make a semi-custom integrated circuit
- use a 50 Ω amplification system using MMICs

The design cost of these solutions is considerably more than for the opamp solutions. You are not talking factors of 10× either. The design cost could easily increase from $50 to $10,000. This emphasises the idea of step changes in amplification costs. If you need to go another 20% higher in bandwidth you can find yourself in a new range and therefore involved in huge cost increases. It is therefore of great importance to make the cheap technology run up to as high a frequency as possible. In this regard, a fixed ×10 amplifier with DC-1.8 GHz performance[‡] for $3 is a good compromise between a low cost opamp and a MMIC.

I am going to give you a simple introduction to the design process by using an amplifying device with a (voltage) gain-bandwidth product of GBW. I am going to say that I can get a bandwidth of GBW at a (voltage) gain of 1; a bandwidth of GBW/10 at a gain of 10 &c. This is a nice simple "text book" mathematical analysis. The resultant amplifier is taken to have a simple single-pole response.

If I want a gain of 100 what bandwidth can I achieve using this amplifying device? Well if I use one, then I get a bandwidth of GBW/100. If I want more bandwidth then I need to spend more money. If I use two stages, each with a gain of 10× then each stage has a bandwidth of GBW/10 and the cascaded bandwidth is GBW/15.5. Or I could spend more money and use three cascaded stages, each with a gain of 4.642 and a bandwidth of GBW/4.642. The resultant bandwidth is now GBW/9.10. You can see the law of diminishing returns here. You get progressively less bandwidth improvement for the extra stages that are added. The tolerance on the overall gain is also increased due to the additional gain uncertainties of each stage.

The bandwidth is conventionally defined as that point where the power output is reduced by a factor of 2 from its LF value. This means that the voltage output is reduced by a factor of $\sqrt{2}$. If there are N identical stages then the output of each individual stage

[†] eg National Semiconductor LMH6624

[‡] Texas Instruments THS4303

will be down by $\sqrt{2^{1/N}}$ at the overall 3 dB point. However, using the simple gain-bandwidth rule, each stage has a bandwidth of $B = \dfrac{GBW}{A^{1/N}}$, where A is the overall LF gain. Using the normalised transfer function of each stage as $T = \dfrac{1}{1 + j \cdot \dfrac{f}{B}}$, the gain magnitudes can be equated: $\dfrac{1}{\sqrt{1+\left(\dfrac{f}{B}\right)^2}} = \dfrac{1}{\sqrt{2^{1/N}}}$ giving $1+\left(\dfrac{f}{B}\right)^2 = 2^{1/N}$ then

$$\frac{f}{B} = \sqrt{2^{1/N} - 1}\,.$$

The overall bandwidth, B' is given by:

$$\boxed{B' = \frac{GBW}{A^{1/N}} \cdot \sqrt{2^{1/N} - 1}}$$

Inspection of this equation shows that when the desired LF gain is higher, the benefit from having more stages is much greater. For a gain of 10 there is little point in using more than two stages. For a gain of 1000, four or five stages are preferable. This equation is pretty simplistic and you wouldn't really use it in practice. It is the idea of splitting the gain between stages that is important.

Although the bandwidth is improved, the tolerance on the low frequency gain gets worse. If each stage gain is set by two ±1% resistors, the worst error increases from ±2% for one stage to ±6% (worst case) for three stages.

17.5 Surviving Component Failure

When a component fails, something must happen to the circuit containing that component. It is obviously desirable that this something is not unnecessarily bad. The first thing you must decide is what possible failure modes there are for any particular component. Then you can evaluate what will happen as a result.

For safety testing it is usual to consider the extremes of open-circuit and short-circuit failure. For other uses, *parametric failure* is also considered. Parametric failure means that one or more aspects of the device's characteristic no longer meet the manufacturer's spec. These include, but are not limited to:

☹ The current gain of a transistor going low.
☹ The resistance of a resistor going outside of its tolerance band.
☹ The insulation of a transformer breaking down at a lower than normal value.
☹ The ESR of an electrolytic capacitor becoming higher than stated.

Question: What *should* happen when the component fails? There are two schools of thought on this which could be called the 'totally dead' and the 'limping' modes respectively. For a car there is no doubt about which mode is preferable. If the alternator {generator} fails, it is better if it goes into a partially working state so that you can at least limp home. Nothing is worse than being stuck out in the middle of nowhere with a dead vehicle.

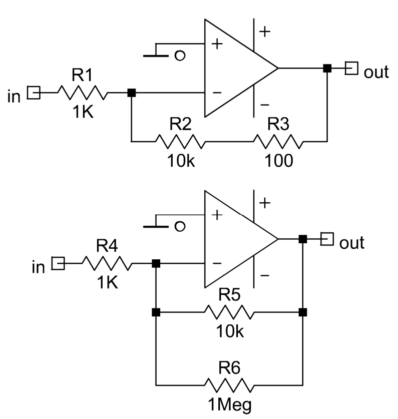
For a piece of test equipment the answer is not so clear-cut {definite}. Let's suppose that you have a simple moving coil meter and a component breaks inside. Would you prefer that the meter gave no reading at all or that it produced a reading that was in error by 10% [and therefore well out of spec]? Well it would be a lot safer in terms of measurement confidence if the meter either read correctly or didn't read at all. That way you would know when the meter had failed. If it was simply reading in error by 10%, you might not realise that it was broken for some time.

You might call this a philosophical dilemma {problem}. You have to decide which design philosophy to follow: should the equipment break totally or limp? This has to depend on the equipment and its use [and the opinion of your boss]. Personally I hate totally dead equipment. If the equipment is going to decide that it is faulty and just shut down then as it gets older it might decide it was faulty all the time and never allow itself to be used. Additionally, the cost of repair might be prohibitive and the equipment would then be completely valueless. I would tend to prefer a limp *and inform* strategy. It should carry on functioning as best as it can, but let the user know that there is a problem with some particular function.

For modern electronic instrumentation this is not a difficult feature to implement. Microprocessor controlled systems can easily have built-in diagnostic and self-checking routines. These should be able to isolate the problem down to a small area of circuitry. For example, on a multi-channel data acquisition system it would be more acceptable to shut down one of 16 channels than to declare that the whole instrument was out of calibration and needed to be returned for repair. This allows the customer {end-user} to decide what to do about the problem.

The argument can be taken to extremes. I have heard the view expressed that the use of series or parallel gain setting resistors makes all the difference to this aspect of the design. For example, here are two opamp stages where a slight trim on the gain has been necessary (in order to use standard preferred values).

FIGURE 17.5A:



In these two schemes the gain is adjusted slightly. Clearly if R3 fails open-circuit the amplifier is dead, whereas if R6 fails open circuit the circuit is only slightly out of calibration. (Short-circuit failure of a small-signal resistor is very unlikely. It is much more likely that the film will crack or the termination will come loose so that the resistor becomes open-circuit.)

I would not differentiate between these two circuits for reliability reasons. There would probably be lots of other places for the signal path to get broken and these would make the effort of differentiating between the two schemes worthless. Both of these circuits can be used.

The parallel scheme reduces the gain, whilst the series scheme increases the gain. You could also use the series or parallel connection at the input to reverse the direction of the trimming effect. As the resistor values get larger, it is more convenient to trim the value with a series resistor. For example, a 1% trim on a 1 MΩ resistor requires a parallel 100 MΩ resistor. It would be easier to use a 10 kΩ series resistor in this case.

Another preference of one scheme over the other is stray capacitance. The series scheme can be used to minimise the capacitance on the summing junction of the opamp, which is always a good thing to do.

17.6 The Hands-On approach

In engineering circles people talk about “hands-on experience” and “getting your hands dirty”. This is not a figurative statement for analog engineers. Hands, and more particularly fingers, are a *vital* part of an analog engineer’s toolkit. Firstly there is the gentle and tentative {hesitant; cautious} running of ones fingers across a new prototype board to see if anything is getting too hot.

That is not quite first in the debug section. First you switch on: you wear safety glasses and put a transparent plastic protector over the board. You can then look for the little tell-tale puffs of smoke which indicate a serious fault. Once the board has been switched on and off a few times it is then relatively safe to approach. You then need to know what circuitry is on the board to avoid touching live circuitry {mains related} or voltages greater than 40 V.

Now when I said fingers, I meant *fingers* and not rings or watches. Don’t wear rings and don’t wear a watch with a metal body or strap when doing this hands-on work. Either take them off or invest in a cheap all-plastic watch.

The reason that fingers are so valuable is speed. If there is a >30 MHz parasitic oscillation on a board, you can usually stop it by touching it, or touching something near it; this is very fast. You can cover hundreds of components in a few tens of seconds. The trouble with >30 MHz oscillations is that they get everywhere and you can measure them everywhere on a board. Just connect your probe somewhere and leave it there, then let your fingers run across the surface of the board actually touching component leads and PCB tracks. You can move the spectrum analyser probe around, trying to get closer to the source by seeing the signal get bigger, but this method of using your hands is often faster. It is also something that you don’t see in ‘academic’ text books.

Another test you can and should do with a prototype is to get a can of freezer spray and just go spraying areas of the PCB to see if anything stops working as a result. If you get ‘over enthusiastic’ with the freezer spray you will get condensation {water} on the PCB and many circuits will stop working as a result. Let the condensation clear and see if the circuit now functions. The point is that the temperature shift will effectively run the component through a range of values, possibly simulating a production shift of tolerances. Doing this sort of ten minute test can save a lot of time. A heating up test with an ordinary domestic hair-dryer should also be done.

Note that these heating and cooling tests also apply to digital circuits and can make a digital system “fall over” {malfunction; break}. Digital systems have three states: working, non-working, and working most of the time with an occasional fault. It is these occasionally failing circuits that are the most difficult to fault-find. In reality a digital system that fails “randomly” with no discernible pattern is probably failing due to some basic fault, like violation of ***setup or hold time*** on latched devices, invalid power rails due to noise, noise on clock lines, bus contentions giving invalid logic levels &c. Heating or cooling the devices in question can shift the timings enough to make the system fail more rapidly and this helps to narrow down the area of the investigation.

There is a lot to be said for ‘getting your hands dirty’ when evaluating components for new designs. Internal construction of relays is of special interest. Whilst you can evaluate relays according to their data sheets, it is possible to use relays for purposes other than the simple applications that they were designed for. It is not unusual, for example, to be using relays as range selection devices in switched-gain amplifiers. The relays may not be specifically designed with this in mind and therefore they may not be specified for your application.

As an example, some high frequency relays, designed for switching >30 MHz signals, are characterised only for 50 Ω operation. If you are not working in a 50 Ω system then this characterisation is not of too much use. It is then valuable to cut the relay open and see how it is made. You can very quickly get an idea of the usefulness of the relay by seeing its internal construction. Some relays have the contact circuit and the magnetic circuit connected together. This can give excessive coil-to-contact capacitance, causing problems for high impedance (>100 kΩ; <30 pF) and high frequency (>30 MHz) circuits. High breakdown voltage (>300 V) is a good starting point for low capacitance; the gaps are larger.

Dissection of components is a valuable thing to do. It enables you to think of specs that you might otherwise not have thought of. Capacitance from coil to contact is one thing that you would expect to measure, but how about capacitance from the contact to the rest of the equipment. That is not something you would immediately think of, but if the contact has a large metal plate attached, it is going to act as an effective antenna to pick up any fields in the vicinity.

When testing a particular part of a circuit it is useful to see the effect of any proposed change immediately. If you have to power down the system, solder a part in place, then power it up again, the change will not be quite as obvious. In this case you may want to resort to live soldering, by which I mean soldering components in place when the power is still applied to the circuit. The soldering iron needs to be isolated to do this, and the circuit being worked on must not be hazardous, both in terms of the voltage (<40 V) and the peak current capability (<1 A). Also the parts should not be so closely packed that you risk damaging the circuitry by shorting with solder.

Another, less risky, technique is to solder in the new component at one end with a little piece of bendy tinned copper wire on the other end. With the circuit powered up you can push the wire with a plastic tool, completing the circuit, and seeing the effect immediately. This gives an instant answer as to the effect of the new component and minimises mistakes.

Capacitors are commercially available attached to 10 cm plastic rods.[†] These are again useful for seeing the effects of a change immediately.

17.7 Second Sourcing

Second sourcing means having a component supplied by at least two different manufacturers. This is an area with no mathematical precision, but with *extreme* opinions. If you don't do what you are told by your superiors then you are wrong, regardless of your own unworthy opinions.

As far as a stable, reproducible design is concerned, you should use parts from only one vendor {supplier; manufacturer}. Every time you have more than one vendor

[†] American Technical Ceramics: ATC Multilayer Capacitor Tuning Sticks ®

supplying a part, you add the possibility that the new part will interact badly with some other part in the system.

As a beginner in the field of volume manufacture you will think that your elders are foolish for being so conservative {cautious}. How can changing a ceramic decoupling capacitor from one manufacturer to another cause any trouble? It just doesn't make sense. Well it does make sense if you know what the problems are. I hope to convince you that using components from multiple manufacturers is a real problem. You still have to do it, but I want to warn you what to look out for.

Before I show you the problems that multiple sourcing can bring, let's deal with the opposite problem; single sourcing. If your design uses components with only one specified supplier for each individual component then the purchasing people are not happy. What happens if the price goes up? What happens if the manufacturer's factory burns down? What happens if there is a trade embargo {import restrictions} from that country? Purchasing people always want as many approved sources as possible.

So there are two conflicting viewpoints that need to be resolved. Do you single-source to ease the design constraints, or multiple source to ease the purchasing constraints? The answer is a compromise; it is somewhere in the middle ground. 'Boring' {no technical interest} commodity components such as low power resistors, capacitors, LF transistors, diodes and logic devices should be multiply sourced. Specialist parts like display devices, transformers, thick and thin-film resistor arrays, custom and semi-custom analog and digital ICs, high speed ADCs and DACs and so on probably cannot be multiply sourced because they are either not compatible with other devices, or you need to pay a large ***NRE*** to get the device tooled in the first place.

Then there are the parts that you have to worry about; the in-between components. It is not certain if the devices are truly compatible and you have to use some judgement as to whether or not to give a second source. Let's take the case of RF transistors. Let's say we are talking about transistors with f_t values of 1 GHz and above. These devices are very difficult to second source reliably. It is actually bad enough dealing with one supplier with their (unspecified) batch-to-batch variation, but when you add in the extra complication of one or more other manufacturers, life in the design department gets rather tough. It doesn't matter one little bit that the device has the exact same type number, spec and package. One *will be different* to the other (unless they are 'badged versions' of the same thing.)

You will only get the feel for this after you have had it happen to you a few times. Then you will be converted into a believer. The ideal situation would be to try the new part in every possible position on every board currently made. This certainly discourages people from changing types too often, but it also reduces the cost because there is considerable engineering expense involved in tracking down weird effects caused by changing component manufacturers. Failing that, make a note of which part(s) have changed and keep on eye on the next batch of units, specifically looking for problems.

With transistors the problems usually include oscillation. One manufacturer's transistor will just oscillate in a particular position in a circuit, whereas the other manufacturer's device doesn't. Perhaps learned professors will sit in their dusty rooms and shake their heads slowly and sadly at this poor state of affairs. Well this is the difference between an academic life and a manufacturing position. If you are busy designing new circuitry (and who isn't) and the production line is stopped because the circuit is oscillating, senior people in the company may start looking unfavourably at the unworthy individual who signed the change form to use the 'defective' part (which now has to be replaced at an exorbitant cost).

The new part may be 'better', having more gain, more f_t, less parasitic base resistance &c, but the bottom line is that it is not compatible with the existing design. Your only course of action is to use as few sources as possible and check all manufacturers' types before committing to volume manufacture. And then, when things go wrong, find a known working product and look for the parts that have changed. It can be that the manufacturer has moved production facilities without telling you. In this case the plastic packages are often marked differently. Sometimes it is down to the date code on the device. *Any visible change* is a key clue in the investigation. [Good production personnel will do this work for you and give you a complete package of fault and reason. This saves an enormous amount of time.]

Semiconductors are notorious for having higher or lower bandwidths or delays than the original working parts, and the production process inevitably finds any weakness in the design. Sometimes the reason for the failure is far from evident. In one case, a relatively low frequency transistor was found to oscillate in only one of a pair of driver circuits. The circuits were nominally identical and they were well away from other circuitry. And yet when there was a problem, it was always with just one of the transistors. The solution was to thread a ferrite bead on the base lead of the troublesome transistor, but it was never clear why one oscillated and the other didn't.

When I said that logic gates could be multiply sourced, I should specifically exclude CMOS switches, multiplexors, and phase-locked-loop chips. These cause no end of trouble when used from multiple manufacturers. At least one manufacturer's CMOS analog multiplexor was found to give huge cross-talk currents when switching a high load on another channel. This wrecked the multi-channel sample and hold circuit that it was used in. The cost of replacing all the chips on all the assemblies running through the factory was far more than could ever have been saved by using a "cheaper" supplier.

17.8 Noise Reduction

By noise here I mean any unwanted disturbance to the signal. I am not referring to harmonic or sub-harmonic distortion however. You do see guidelines for reducing noise given in some publications. They contain general principles like:

- keep analog and digital grounds separate except at one point
- keep the decoupling capacitors close to the chips
- have lots of decoupling capacitors
- use different values of decoupling capacitors in parallel
- screen sensitive circuitry
- keep switched-mode power supplies in separate areas.

These rules are useful to make a start on a design, but they do not give an understanding of the noise reduction process. Noise reduction is perhaps one of the key areas of 'black art' that separate the good engineer from the mediocre {average} one.

It would be foolish to think that I could condense noise debugging down to a few concise pages of text. Nevertheless I am going to try to present some aspects of the process in a way that can be followed without requiring too many flashes of insight.

There are two very different sorts of noise that you will be concerned with. Firstly there is the interference generated by other circuitry, either within the equipment being designed or from the outside world. Some authors prefer to call this noise by its correct name, *interference*, and I have no objection to that. Secondly there is noise within the devices used, which may be Gaussian, or 1/f, or other types, and which is very much a function of the initial 'paper' design of the system; this I will call *inherent noise*.

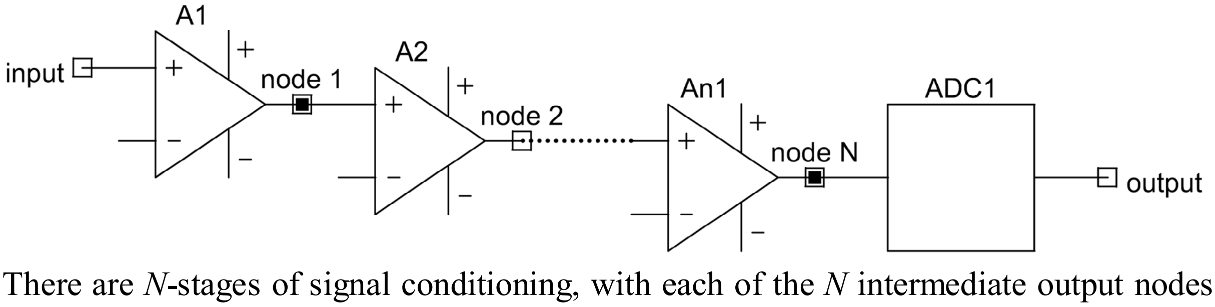
The first thing I need is a system to fix. A generalised analog sub-system can be represented by an input, a series of one or more signal conditioners, and an output device. Consider, for example, a domestic audio amplifier system {hi-fi; stereo}. There would be an input from say a tape deck, the signal conditioning would be an amplifier chain, and the output would be a power amplifier to the loudspeakers.

For a DVM the input would be the front panel terminals. The signal conditioning would be the amplifiers, and the output device would be the ADC, as far as the analog system is concerned. After the ADC the data is digital, so noise problems automatically stop at that point. Data integrity problems do cause noise, but that is a whole other subject.

Obviously the signal conditioning could be more complex than simple amplification. It might be a voltage-to-frequency conversion; it might be a logarithmic compression. Whatever it is there will be a voltage generated that you can measure. *Measurement* and *test* are the key words here. You can either scatter capacitors around the board and hope that they do something, or you can make measurements and tests to see what needs to be done.

Remember that you may be picking up somebody else's design and you are being asked to 'make it work'. It is all too easy to look at the design and think that it is rubbish because that is not the way you would have done it. However, you have to realise that *almost any design can be made to work if you try hard enough*. It is just that you may have to try so hard that it becomes uneconomic, so you have to know when to scrap a particular approach and try a different one. All I am saying is just don't give up too soon. In any case, here is the model of the analog system I have chosen to debug.

FIGURE 17.8A:



There are *N*-stages of signal conditioning, with each of the *N* intermediate output nodes accessible. The overall output may be either digital or analog. Let's say by way of illustration that the output device is an ADC and the signal conditioning blocks are just amplifiers. This makes the system less abstract.

The very first thing you have to realise about the noise reduction phase of a design is that there is *never* just one noise source. Don't expect to kill all the noise in one go with a capacitor 'in the right place'. You have to locate noise sources one at a time and kill off each one until the noise is acceptable.

Inherent Noise

Inherent means 'belonging naturally to' or 'being a quality of'. This would be the first part of the design if you are starting from a blank sheet of paper. Regardless of interference sources, the basic design has to work to the desired noise level. If you get this part wrong then no amount of ‘noise debug’ work will fix it. To evaluate the inherent noise of the system theoretically you need to know several things:

- □ What is the required system bandwidth to the output device?
- □ What is source impedance of the input device?
- □ Is the source impedance constant over the system bandwidth?
- □ What is the inherent noise of the input device?
- □ What is the specified requirement for RMS noise at the input terminals of the output device?

This is the ‘paper design’. The first task is to get the design to work on paper; perhaps the more modern phraseology should be that the design works on the simulator!

If you are making a general piece of test equipment, you will not have the details of the source. You will then have an even more complicated task because you need to minimise the noise over a band of input impedances. Some assumptions about the source will have to be made.

For now let’s keep it simple and say that you have a given bandwidth B from 3 stages of equal voltage gain, A. If you have the luxury of being able to limit the bandwidth as a simple single-pole low-pass filter at any point in the system, where should that be done; at the input, at the output or should it be distributed? Think about that before carrying on.

To give me some figures to work with, I am going to inject a sinusoidal signal at the front of the amplifier E_n. This is going to be at some fixed frequency, which I may change as I see fit to investigate the performance of the system. The first thing to say is that for signals below the bandwidth of the system, the signal E_n is amplified by A^3 and it is unaffected by the positioning of any filter(s). Hence for this part of the analysis you are only concerned with signals *above* the system bandwidth.

You might be surprised by that statement. Surely you can ignore signals above the system bandwidth? No! No system has a ***brickwall*** filter response. Noise is summed to infinity in frequency terms, so the slope and position of the filter characteristics can be important. I am going to make the simplifying assumption that the gain of the basic amplifier blocks is A/2 at a frequency of 2·B.

If the bandlimit is placed after the first stage you get a voltage gain to the noise at a frequency of 2·B of $\frac{A}{2} \cdot A^2$. If the bandlimit is placed at the output of the last stage you get a gain to the noise signal of $A^2 \cdot \frac{A}{2}$, which is identically equal to the previous value. Thus for an *external* noise source, it does not matter where in the chain of amplifiers the bandwidth is limited.

If one of the amplifier stages is itself a large source of any sort of noise, filtering is best done *after* that stage. There is however one exception to this rule, and that relates to the input. You may find that a signal frequency can be applied that exceeds the capability of the amplifiers to behave linearly. In this case it is important to filter this signal out before it reaches the amplifier. A signal which is well above the bandwidth of an amplifier can cause non-linear effects such as rectification or ***intermodulation***.

Let me give you an example of that problem. If you have a DVM with a 10 Hz input bandwidth and you attach long wires to the input you will find that the wires will pick up the ambient RF field. This field includes frequencies from 50 Hz right up through the GHz region. An amplifier designed to work at 10 Hz may have trouble with these higher frequencies. What you will find in practice is that when the RF power is higher, the DC reading is different. This is a very bad situation for precision measurements as you have an uncontrolled 'random' measurement error. There may be no indication on the DVM that the reading is wrong. A radio signal may be modulated and this modulation, when rectified may show up as noise, but an interfering signal may not be amplitude modulated and the resulting rectified signal may not cause noise but only a small DC offset. For this reason DVMs and scopes traditionally have a small input resistor and/or capacitor right at the very input to filter these out-of-band signals and give less chance of rectification of the received RF fields. The user should also employ screened and/or twisted test leads in the measurement setup in order to minimise the exposure to such RF pickup.

Let's get back to the inherent noise. I said that the noise is summed to infinity and that this may be important. I should quantify that. To do so I will make a model of a system. In this model the noise is all assumed to be at the input of the first amplifier in the chain and there is a single-pole RC filter at the input to the output device (which does not have any internal filtering of its own).

EX 17.8.1: If the input noise density has a flat frequency spectrum from DC to daylight at a level of E_n (nV/√Hz)

a) What is the total RMS noise at the input to the output device when the system bandwidth is equivalent to a single-pole filter of bandwidth B?
b) What is the error if the noise is assumed flat up to the bandlimit point and then stops completely?

Having as many poles as possible in the measurement system minimises the noise. Realistically, however, only two or three poles are needed to give most of the available improvement. Also, the filtering needs to be more towards the output end of the amplifier chain in order to minimise the noise. Again this is not too important provided the first stage has some gain.

***EX 17.8.2:** A three stage amplifier has voltage gains of A_1, A_2 and A_3, with A_1 being the first stage. The RMS voltage noises, referred to the input of each individual stage are given by E_{N1}, E_{N2} and E_{N3} respectively. Neglect the current noise.

a) What is the noise voltage at the overall output of the amplifier?
b) What is the noise voltage at the output of the amplifier, *referred to the input*?

Now the reason that last question got you to refer the noise to the input is because you want to know how much noise you are going to put onto the input signal. It tells you the measurement limit that you can achieve. For example, a particular DVM might have a noise referred to its input of 125 nV RMS. Such an instrument would clearly be unsuitable for measuring voltages around the tens of nanovolt region.

Actually I can generalise the answer to that last exercise/example by making a simplifying proposition. I can break the amplifier into two parts, the front and back parts.

I can obviously do this on an amplifier having any arbitrary number of stages greater than or equal to 2. Using the same notation as before the noise referred to input is just a simplification of the answer to that exercise/example. Namely:

$$\boxed{E_{N[input]} = \sqrt{E_{N1}^2 + \frac{E_{N2}^2}{A_1^2}} = E_{N1} \cdot \sqrt{1 + \left(\frac{E_{N2}}{A_1 \cdot E_{N1}} \right)^2} = E_{N1} \cdot \sqrt{1 + \delta^2}}$$

If the factor δ is smaller than 0.458, the output stage will not increase the overall noise by more than 10%. In other words, you can neglect all but the front part of the amplifier if $A_1 \times E_{N1} > 2.2 \times E_{N2}$. *This is a very useful result.*

There is actually a bit more to be said about this simplified system. It may be that the voltage gain is fixed by the type of amplifier block that is used. You may have a choice of an amplifier with a lower voltage gain and a lower input noise voltage. Which is a better amplifier to use?

EX 17.8.3: You have two amplifier modules to wire in cascade. One has an input noise of E_{N1} and a voltage gain of A_1. The other has an input noise of E_{N2} and a voltage gain of A_2. Which should be closest to the input in order to minimise the overall noise?

If any stage is driven from a significant source impedance, the current noise of the amplifier should be multiplied by the source resistance to give an additional amount of input-referred voltage noise. The noise values for that stage are then combined using the RSS method: $E'_N = \sqrt{E_N^2 + I_N^2 \cdot R_S^2}$

The exercise question above is somewhat 'academic', as presented, since the amplifier block would probably be an opamp. Given that an opamp has 'infinite gain' you may consider the preceding exercise irrelevant. However, further thought is required. Opamps have large DC gains, but the gain-bandwidth products are far from infinite. Thus in order to achieve a given bandwidth you may have to run a particular opamp at a lower closed-loop gain than another opamp. In this case you are back to the situation given in the exercise.

As there are hundreds of opamps to choose from, you are always required to choose the best one for your application. For non-critical applications, where any device will do, choose the cheapest, or perhaps a type already used on that PCB assembly, or whatever is most readily available. There is no one criterion which defines "the best".

Often you will want to pick the least noisy opamp that you can afford. It may be that the absolute lowest noise device costs 5× the price of one which produces only 10% more noise. In this case you will have to decide if the slightly increased noise is acceptable, given the cost reduction. Suppose the cheap opamp is \$0.50 and the expensive one is \$10, but the noise is half as much on the expensive one. For the front end of a \$9000 piece of equipment this extra expense may be a good investment.

For now let me just evaluate the noise performance of the various opamps to give you a numerical basis on which to see which opamp is quietest. Now manufacturers' data sheets are trying to get you to buy their parts. When they say the voltage noise of the part is 3 nV/√Hz you would be forgiven for thinking that this therefore makes the opamp quieter than another part which has a voltage noise spec of 7 nV/√Hz. You must also check the *corner frequency* of the noise. Typically the voltage noise of an opamp is constant with frequency over a wide range of high frequencies, but as the frequency is decreased the noise suddenly starts to ramp up at 10 dB/decade. What was 3 nV/√Hz at 1 kHz may then turn into 10 nV/√Hz at 100 Hz.

Remembering that the noise values given are RMS quantities, and that noise *power* adds, the noise is evaluated by taking the square root of the integral of the squared noise voltage. The equation used for the noise as a function of frequency is:

$E_n = E_{HF} \cdot \sqrt{1 + \left(\frac{f_C}{f}\right)}$ where E_{HF} is the high frequency asymptote of the voltage noise density, measured in nV/√Hz, and f_C is the voltage noise $1/f$ corner frequency.

$$\boxed{V_{RMS}^2 = \int_{f_L}^{f_U} E_{HF}^2 \cdot \left(1 + \frac{f_C}{f}\right) \cdot df = E_{HF}^2 \left[(f_U - f_L) + f_C \cdot \ln\left(\frac{f_U}{f_L}\right)\right]}$$

EX 17.8.4: Opamp A has 2 nV/√Hz and a $1/f$ noise corner of 300 Hz. Opamp B has 3 nV/√Hz and a noise corner of 70 Hz. Which opamp has the least voltage noise over the range 1 Hz to 1 kHz?

The noise situation is further complicated by the current noise. The current noise also has a $1/f$ corner frequency, but this will not be the same as the voltage noise $1/f$ corner frequency. The current noise summation formula is the same as the voltage noise summation formula above, but with current noise substituted for voltage noise.

A simplification can be made when considering the combined effect of current and voltage noise. Consider a low-noise bipolar opamp with 2 nV/√Hz and 20 pA/√Hz. The current noise will equal the voltage noise when the source resistance is 100 Ω, which I will refer to as the *noise cross-over resistance,* $R_{NC} \equiv \frac{E_n}{I_n}$

Now don't make the mistake of thinking that the noise has magically been optimised {minimised} by making the current noise and the voltage noise contributions equal. The overall noise is always least when the resistance is zero. What can be said, however, is that this cross-over resistance is a convenient calculation short-cut, since the voltage noise and current noise contributions are combined by squaring, adding, then square rooting. If one factor is 3× larger than the other, the smaller factor can be neglected and the resulting error is less than 5%. Hence for source impedances less than one third of the noise cross-over resistance, the current noise can be neglected. Likewise for source resistances larger than three times the noise cross-over resistance, the voltage noise can be neglected.

Typically low-noise bipolar opamps give the lowest voltage noise, but have low noise cross-over resistances (<1 kΩ). FET and CMOS opamps have higher voltage noise, but high noise cross-over resistances (>100 kΩ).

AD811	bipolar	1.9 nV/√Hz	20 pA/√Hz	$R_{NC} = 95\,\Omega$
AD8610	JFET	6.0 nV/√Hz	5 fA/√Hz	$R_{NC} = 1.2\,\text{M}\Omega$

To see which opamp is quieter in the region 300 Ω to 400 kΩ it is easy to see that the AD811 is current noise dominated and the AD8610 is voltage noise dominated. The cross-over point between these two opamps then occurs when the source resistance is equal to 6 nV/20 pA = 300 Ω. The apparently noisy JFET device is therefore the quieter opamp in any application having a source impedance above 300 Ω.

System Interference

System interference cannot be calculated at the stage of the paper design. You can look at the power supply rejection of the amplifiers and get figures for the power supply ripple, but those simple calculations are not ordinarily the things that catch you out.

If your design doesn't work due to excess 'noise', you can just stare at the circuit diagram for a long time without coming up with reasons for the problem. The reason for this failure is that *the problem isn't explicitly drawn on the circuit diagram*.

All that you know is that there is too much noise on the output, and it is your task to fix it. Now there are two basic approaches:

☺ Start at the input and measure at the intermediate nodes on the way towards node N looking for noise.

☺ Start at node N and disconnect the preceding stages to see which of them is generating the noise.

If you have a large multi-stage circuit you might even do a *binary search* for a specific noise component, using either method, by starting at node *N*/2 &c. In any case, the key thing to know about noise debug is that you have to measure to a very low noise level at each stage in order to find all the noise sources. That needs more explanation. Let's suppose that the output device is a multi-digit digital display. The measure of the noise on the display is the max-min reading on the display over a 10 second interval, for example. If you want this value down to say less than 2 digits and it is currently at 20 digits there is no use in leaving the system intact and trying decoupling capacitors and various other things to see if the noise can be reduced. You will not see a reduction unless you are very lucky. By very lucky I mean that you both guess the right location of the component to add/adjust amongst all the values on the board, and that this noise source is *dominant*.

This is why I said that the noise level has to be low when you are doing noise hunting. You need to be able to see a noise level difference of less than 1 least significant digit on such a system, and the only way to do this is to have very low noise to start with. Actually it is good to use the output device as the monitoring device. This way you *know* that you do not have a measurement problem. In the example circuit given, the first thing I would do would be to open-circuit the output of amplifier A_n and link the ADC input to the ADC reference midpoint, or to some heavily decoupled passive voltage source. This is the first step on the second noise debug method.

If you can't get a quiet reading on the display now, then you know that you have work to do on the ADC block. This has limited the range over which to hunt for the problem. The procedure for fixing the noise comes later. For now let's just find the area that the problem lies in.

If there was no problem with the ADC block on its own, then you reconnect the output of A_n and go back a stage to A_{n-1}. Again you open-circuit the forward path and short-circuit the input to the next (forward) stage (A_n in this case). This probably seems very straightforward, and it is. The thing is to know what to do before you are confronted with a noise problem. One thing is for sure; the final noise problems will only be found when the system is virtually complete and otherwise nearly ready to be shipped. You will therefore be under greater than usual time pressure to get the noise fixed.

As a matter of practical experience on projects, the whole project will be running to schedule on each individual sub-assembly. Then, when you put them all together, the system is a mess. Even the simple act of screwing on the covers can cause an inordinate increase in noise, just when you thought the product was ready to ship out!

There is one additional thing to know about the short-circuiting of the input to the next stage. If the stages have relatively high output impedances, let's say above a few hundred ohms, then the next stage should be fed from a similar valued resistor shorted to ground or a quiet DC bias point. This handles any impedance related noise problem. This technique is then continued back through the amplifier chain until the noise is located and eliminated.

There is a problem with this noise debug method, however. It can be that you never get to see the frequency, repetition rate or waveshape of the noise waveform you are trying to eliminate. Actually seeing the waveform on a scope is a very good way of tracking it down. For example, if the waveform is at the frequency of the mains line voltage, you may suspect either direct magnetic or electric field interaction, or perhaps ripple from a half-wave rectified supply. Double mains frequency interference would be due to full-wave rectified supply lines, for example.

For this reason you may want to connect a probe of some description right at the input to the ADC to see what is going on there as you progressively link in more of the circuitry. This probe could be connected to a scope or spectrum analyser, depending on what type of system you are dealing with. Just remember one vital point: Measure the system noise on the output device *before* and *after* the connection of this additional probe. It is not at all unusual for the probe itself to inject/cause additional noise!

The other debug method mentioned was to start from the beginning of the amplifier and work your way back to the output device. This is also a sensible technique because, as mentioned earlier, most of the noise is going to be generated by the earlier stages in the amplifier chain.

Again you will want to remove the input device and replace it by its equivalent source impedance, giving an accurate representation of the noise sources. The next thing you need to do is probe the output of the first stage. This is the hardest part of the circuit to probe because the signal level is probably very low. There has been very little gain added, so the full scale signal level can be anywhere from microvolts to tens of millivolts depending on the application. If the signal level is at hundreds of millivolts then it is unlikely that you would be having a serious noise problem! It is difficult to give you a general solution at this point. The probing of a 10 Hz bandwidth circuit and a 100 MHz circuit are somewhat different.

The thing is that there is no signal applied at the moment, so you are interested in the noise level in terms of mV rather than as a percentage of the signal level. Let's suppose that the full scale signal is 50 mV ptp and you are looking for less than 1 LSB of noise on an 8-bit system. 1 LSB is 50 mV/256 = 195 μV. You are not going to see that level on a 5 mV/div scope with a 10:1 probe! It is *vital* that you do this sort of calculation before probing a signal and declaring that you "can't see any noise on it". To do otherwise would be rather like a small child covering up their eyes with their hands and declaring that they can't see anything.

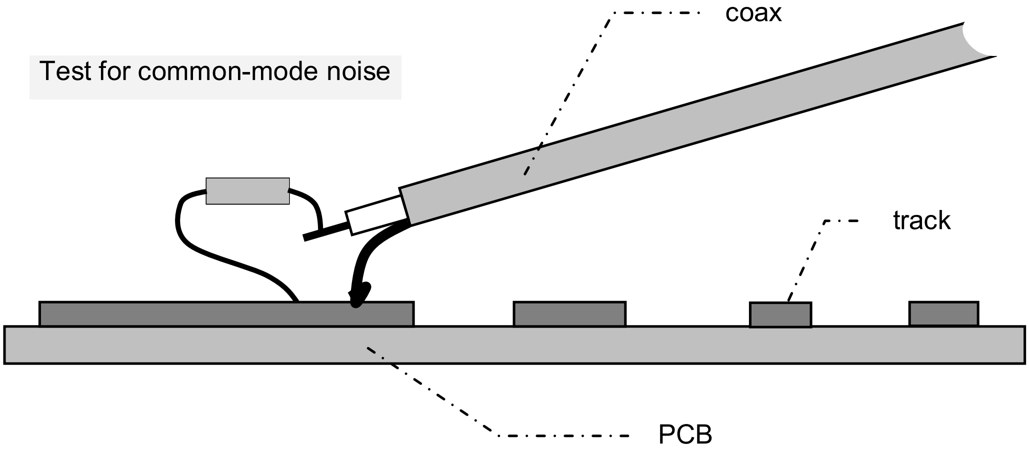
In this particular case, if you use a 2 mV/div scope and a 1:1 (cable) probe then you will have *just* enough resolution to see something. This is where you need to limit the amount of noise seen. Turn the bandlimit on your scope down to a low setting like 1 MHz if possible. This will reduce its own internal noise as well as reducing the noise you are measuring. Then use a piece of coax cable as the probe. Solder the braid to a nearby signal ground in your Equipment Under Test (EUT). If your scope is grounded (this is the normal situation) then make sure that it is acceptable to ground the EUT. If not then use an isolation transformer on the EUT.

Soldering the cable to the signal ground a short distance from the measuring point, < 2 cm, will give you the lowest possible common-mode noise. In fact extra grounds from the scope chassis (or unused BNCs) to other points on the EUT will provide additional common-mode current paths which may further reduce the common-mode noise.

To reduce the normal mode noise, put a wire-ended resistor in series with the coax cable. This should be at least 1 kΩ and could be as much as 100 kΩ. You are trying to deliberately reduce the measurement bandwidth by using the ≈100 pF capacitance of the coax cable as a low-pass filter. If you go higher than 100 kΩ then you will start getting significant measurement errors due to the 1 MΩ input impedance of the scope. (100 kΩ into 1 MΩ gives a 9% gain loss which is not that important when doing noise debug work.) If you need or want more bandwidth reduction, adding some more capacitance at the input to the coax cable may be better than increasing the resistor any further.

There is a quick check that you *must* do at this point. Link the input resistor to the same signal ground that the coax cable screen is soldered to. This tells you if there is a problem with common-mode noise. If the resulting noise is still larger than the noise you are trying to measure, you may have to route the coax cable away from strong fields, or add in extra parallel ground paths to reduce the common-mode current flowing in the coax cable screen.

FIGURE 17.8B:



Using this method you can measure low frequency noise sources such as switched-mode power supplies, display electronics, motor current surges, fan current ripple &c. If, having tried this all the way through, you have still found nothing, then you may need to increase the measurement bandwidth.

A spectrum analyser is a very good way of getting more bandwidth and more sensitivity at the same time. Whilst an ordinary scope will have a highest sensitivity of perhaps 1 mV/div, a spectrum analyser will have a lowest sensitivity of more like −100 dBm, depending on the *resolution bandwidth* used.

EX 17.8.5: 0 dBm in a 50 Ω system means 1 mW in 50 Ω. dBm means dB relative to 1 mW. Express −100 dBm as a ptp voltage for a sinusoidal signal.

There are two essential things to be careful of when using the spectrum analyser: Firstly, the spectrum analyser may have a low frequency limit which is not low enough for your application. I have used an 18 GHz spectrum analyser which had a low frequency limit of 0.01 GHz; that's 10 MHz! Hence you need to check the front panel of the spectrum analyser and/or read the manual.

Secondly, it is essential to avoid 'breaking' the spectrum analyser's front end. The 50 Ω input will be susceptible to DC voltage overload and to static discharge. Some manuals even warn you to discharge coax cables before connecting them to the input in case the cable happens to have acquired a static charge.

Realistically the spectrum analyser is best used to look for processor/digital noise and parasitic oscillations. These will not show up well on a scope. The spectrum analyser should be connected via a coax cable with a 100 nF ceramic capacitor and a 100 Ω resistor both in series with the coax signal conductor at the EUT end of the cable. The AC coupling will protect the spectrum analyser and the 150 Ω load is hopefully acceptable for the EUT.

You can get higher sensitivity scopes (with low bandwidth), or you can buy a small preamp box to boost the gain into the scope; these may give a better answer than the spectrum analyser below 10 MHz. The use of FFT analysis of the data further enhances the measurement resolution, as discussed previously.

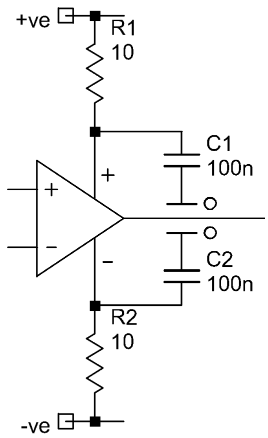
Another trick when hunting low frequency noise (< 1 MHz) is to temporarily boost the gain of the stage you are working on. This might be done by changing resistor values, for example. If the gain is boosted by a factor of 10×, the noise will be much easier to measure. Boosting the gain will undoubtedly reduce the bandwidth, but that may not be a problem for the low frequency interfering signals being looked at.

Interference Reduction

Once you have located the rough area of the problem it is time to actually deal with the source. This will obviously depend on the nature of the interference mechanism and it is impossible to draw this procedure out as a flow chart. This is where you need to follow an intuitive procedure. However, this does not mean that you are now on your own. There are some specific ways that interference can find its way into a circuit and I will take these up one at a time.

The most obvious interference mechanism is due to finite power supply rejection. In other words, noise on the power rails gets onto the amplifier output, or signal conditioner output. This can happen in one of two specific ways: Firstly the device itself will have a defined Power Supply Rejection Ratio [***PSRR***]. For example, an opamp will usually have a specified PSRR and you will find typical curves of how this changes with frequency. Above a corner frequency, the PSRR usually drops at 20 dB/decade; in other words the device is more susceptible to higher frequency interference sources.

FIGURE 17.8C:



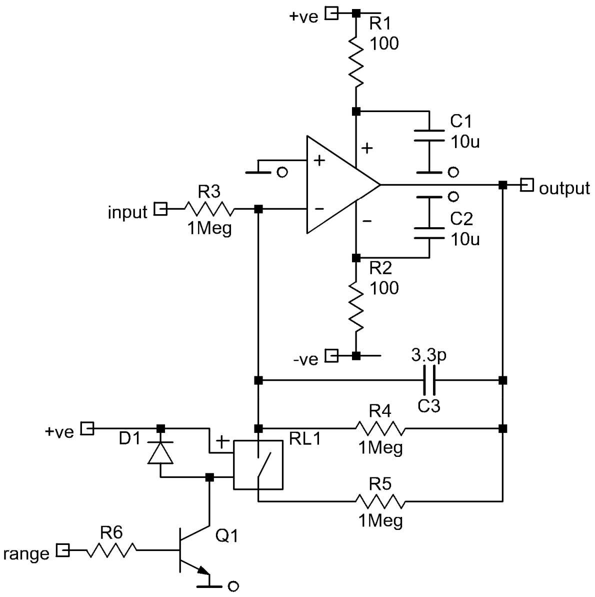
For this reason it is usual to see opamps with their power rails decoupled as shown. (The amplifier input and output connections have been left off so that the power rails can be emphasised.) R1 and R2 might be up to 100 Ω or more, depending on how much current is required from the opamp output. The capacitors would not normally be smaller than 1 nF, with an upper limit of 10 μF.

R1 and R2 provide an impedance for the capacitors to 'work against' {to work *with* really}. If the resistors are left out then the capacitor has only a small amount of track inductance with which to form a low-pass filter. This is not enough for typical interfering frequencies in the range of 10 kHz to 1 MHz. For frequencies above 10 MHz a small ceramic capacitor (10 pF to 1000 pF) will work on its own, but there can be problems with the relatively high Q of such a filter. A small amount of resistance, even as low as 1 Ω can reduce or eliminate this problem. Note that even 100 Hz repetition rate sources can cause noise problems. The key thing to watch out for is the edge speed of this low repetition rate signal.

The power supply rejection ratio of the opamp itself drops at 20 dB/decade. The power supply rejection due to these little filters increases at 20 dB/decade. The overall power supply rejection ratio is therefore maintained at a relatively constant and high value by the use of these simple inexpensive components. The decoupling capacitors can never be included within the opamp itself, the maximum possible capacitance being around 50 pF due to the chip area required.

Application notes consistently omit the resistors R1 and R2 shown above. One can only suppose that this omission is due to the marketing department trying to claim that this opamp can be made to work with minimal external components. Always expect to need these 'extra' resistors and only omit them with reluctance.

FIGURE 17.8D:



The second problem with power supply rejection is also calculable, but only from a knowledge of the parasitics of the components. It is related to the impedance of the circuit.

Here is a typical type of problem. You have identified this stage as the source of 101 kHz noise at a level of 195 μV ptp. RL1 is a reed-relay. You have measured the power rails and they have 133 mV ptp (+ve rail) and 107 mV ptp (–ve rail) of 100 Hz ripple.

Typical actions might include:

- ☹ Look at the circuit diagram for minutes or hours trying to see the problem.
- ☹ Look worried.
- ☹ Shunt C1 and C2 by 10,000 μF caps to *really* make sure.
- ☹ Scratch your head.
- ☹ Short the input to ground.
- ☹ Assume that the opamp is faulty and change it.
- ☹ Change the opamp to a different type.
- ☹ Get a big sheet of copper clad board and individually connect each of the ground points to it to get a solid ground.
- ☹ Drink coffee.
- ☹ Tell your boss it is a difficult problem.
- ☹ Suggest that the problem is due to a PCB problem and that you will need to re-lay the PCB.
- ☹ Actually re-lay the circuit on a 6 layer PCB using a solid ground plane on two layers.

Let's suppose that none of these fix the problem; a very likely situation!

***EX 17.8.6:** What would *you* check or do next?

There is not a set procedure to follow when you have reduced the noise source down to a single stage, other than perhaps to reduce it down a bit more. You ideally need to find something that makes the noise come and go at will. You can then deduce both the *noise source* and the *coupling mechanism* in order to work out a solution. Once the noise source and the coupling mechanism have been established, a definite path to a solution has been established.

Re-laying the PCB on a 'hunch' is not a good solution. You have to understand the problem, even if it is because there is a PCB layout problem. You should model any PCB change by cutting tracks or lifting components, if possible, before doing a re-layout. Otherwise you will be late and over-budget, and still not have a working system. This is not a route to success.

To solve the problem with this particular circuit, you might consider any of the following in no specific order:

- ➢ Remove R5.
- ➢ Remove RL1.
- ➢ Power the circuit from a completely separate supply source such as batteries or (linear) bench power supplies.
- ➢ Wave a 10:1 probe around (connected to a scope) and see if there is any strong electric field that could be interfering with the circuit.
- ➢ Get a coil of some description as a pickup coil connected to a scope and wave it around near the circuit to see if there is any magnetic field that could be interfering with the circuit.
- ➢ Try a shielding plate over or around the circuit.

- Try some mumetal or radio-metal sheet around the circuit to see if that improves things.
- Try shutting down possible interfering sources that might be producing such a frequency of noise. This would include switched-mode power supplies, display backlights, deflection circuitry for CRTs (Cathode Ray Tubes). high current drive circuitry &c.

If you have done all of these steps and the problem persists, then I would think that you had not done one of the steps correctly. For the sake of solving this particular puzzle I will say that removing RL1 removes the 101 kHz noise completely, but introduces a worse (bigger) noise problem at 100 Hz. (The noise increases to 305 μV ptp.) You check for local magnetic and electric fields and none can be detected.

***EX 17.8.7:** This is a very real sort of problem and it is much worse when you are uncertain of your basics. The question is: what is the nature of the problem you have here and how are you going to fix it?

Cancellation is the most puzzling of phenomena when encountered on the bench, and yet as far as equivalent circuits and circuit theory is concerned, cancellation is one of the simplest to understand. You decouple something 'better' and the noise gets worse! In fact if you do something that you think should give a definite improvement, but actually makes the noise worse, you should wake up and pay more attention. If you want a low and reproducible (unit to unit) noise level you must not rely on an unspecified cancellation effect. If you are aware of the mechanism and it is under your control then that is fair enough. However, if you just observe that removing a decoupling cap makes the noise better you must investigate diligently. Either the capacitor was connected to a noisy ground point or the effect is due to cancellation. You need to establish which it is and deal with it appropriately.

I suppose I should say that it is not entirely necessary to understand what the interference source is, or what the coupling mechanism is, in order to eliminate a particular noise signal. You may get lucky. On the other hand, when your luck fails you, or you have multiple noise sources and coupling mechanisms, the systematic methods given here will be invaluable.

When you are trying to eliminate system noise from a circuit you should (ideally) identify where it is coming from. You can and should measure the possible sources of the problem such as power rails. Suppose you have 10 mV ripple on the positive power rail. Is that enough to cause the problem you are seeing? You can calculate that it shouldn't be, but your calculation may not involve some unforeseen interaction, parasitic or stray element. The best way to *prove* that this rail isn't causing the problem is to ***slug*** it. Make the power rail noise reduce by a factor of 10× (or 2× if that is the best you can achieve) and see if it has *any* effect on the output of the amplifier, stage or whatever.

One trick is to have a box of huge capacitors on your bench. If the rail normally has 10 μF decoupling capacitors then tacking a 1000 μF in parallel should kill off the signal [up to say 500 kHz] If it already has 1000 μF capacitors then use a 10,000 μF. Measure the rail to *prove* the ripple has reduced and then see if the noise changes on the output that you are looking at. This is a fast and effective way of narrowing down the search for noise. Obviously the capacitors you will be using for this will be electrolytics, so be careful to connect them up the right way round.

If you tack solder one end of the capacitor to the board and use the other end to make momentary contact on the power rail, then make sure the capacitor can is not pointing at you (or anyone else) in case you connect it incorrectly and it goes bang.

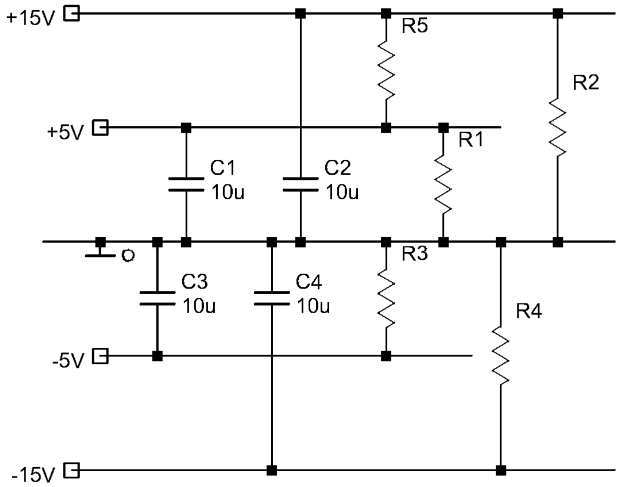
It is not a good idea to be tapping 10,000 μF capacitors onto power rails. They will take very large surge currents and it is not very safe. It is better to turn off the power, wire them in circuit, then switch the power back on. [I always look away from the circuit when I switch on the power, just in case I was distracted when wiring up the circuit and got the capacitor wired-in backwards.] Since you are going to buy or obtain some electrolytics specifically for this purpose, get the best high frequency, low ESR, large capacitance types that you can get. It doesn't matter that they don't fit on the circuit board at this stage. You must first find the source of the problem, then you can work out a suitable remedial solution {fix}.

If you already have 10,000 μF capacitors on the power rails then you could find some bigger ones to put in parallel, but a better solution is to use an external linear power supply. This is where you will be happy that somebody routed all the power through one connection. You will be able to remove the ferrite bead, fuse, or connector pin, feeding the external power supply in at that point.

Problem: how do you stop the circuit doing horrible things at power up? This is a common problem. At power-up in a multi-rail system [even two power rails counts as a multi-rail system] it is often 'unpleasant' if one rail is late coming up. Hence if you switch on the system and then switch on the bench power supply you can blow up the circuit. Likewise, if you apply power via the bench supply before switching on the system power you can get a bang. [It may break quietly without going bang, but the end result is a broken circuit.] People "handle" this problem by switching the on/off switches of the system and the bench supply at the same time, or perhaps from a common power switch. Realise that you will never synchronise the supplies to better than a few tens of milliseconds by this method and it may cause the circuit to fail. However, if this "simultaneous switching" method works in your application then that's fine.

First you need to see the problem. Then you can see how it can be dealt with. Remember, all problems look simple when they are laid out for you like this. When all you know is that the circuit 'blows up' at switch-on, things look much more difficult!

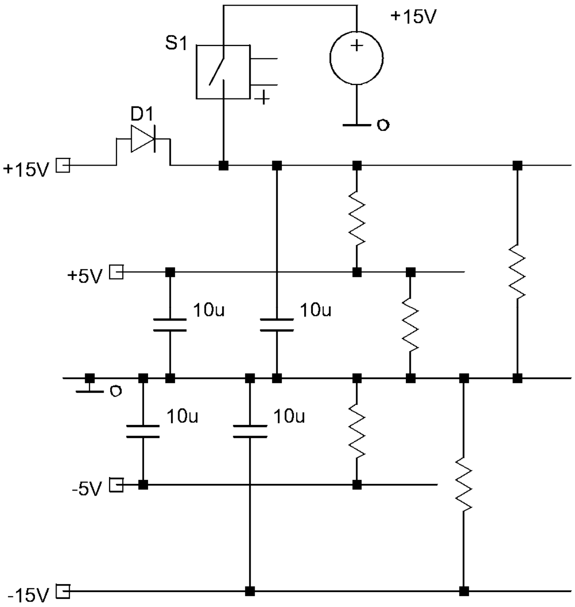
FIGURE 17.8E:



This simplified circuit represents the loads on the power rails. These could be opamps, logic devices, transistorised circuitry &c. In reality, then, the loads will not be linear. Nevertheless this simple model gives a feel for the problem, without hiding the basic ideas.

In this circuit, if R2 is a heavy load (low resistance) the +15 V rail may rise more slowly than the +5 V rail. This can mean that R5 gets reverse voltage applied to it at power up. If it is an integrated circuit, this may be enough to destroy it. This start-up problem can be improved by using diodes between the power rails to limit the reverse voltage to less than about 0.8 V. Schottky diodes can improve on this, but the real answer is to improve the power supply itself.

FIGURE 17.8F:



Here is a solution to the power-on problem, for when you want to feed in power from an external linear (=quiet) bench supply. This might be done, for example, to see if some system noise is coming from a particular power rail. *Don't use a switched-mode bench supply* or you will have a nagging doubt that this may be causing noise of its own. You want positive facts: example, the new linear bench supply is totally quiet and yet there is still noise, therefore the existing power supply scheme is not the source of the noise.

The power-on condition is handled by D1. When the system is powered up, you can then switch on the bench power supply output switch (S1; which is not required to be electronically switched) if it has one, or connect the wire up to the bench supply. This can be done at your leisure. To turn off, make sure that the bench supply is turned off first.

A word of warning. If the noise goes *up* as a result of using the quiet power supply then the power rail **is** contributing noise. The noise is cancelling out with another source and you must eliminate both to get a quiet and repeatable system. [I am assuming that you are measuring the noise voltage on the rail, and that by putting on the external supply you have established that the noise on the rail went down significantly.]

At first, when you are hunting down noise, you have a lot of uncertainties. Is it that stage, or that one, or that one? You must reduce this uncertainty by introducing *certainties*. The more certain you are of the results of any particular test, the more the problem is simplified. Let me give you an actual example. I was hunting down noise on a video display card. The display had been wobbly and fuzzy, and doing all sorts of unpleasant things. I had decoupled the power rails extensively, split them up so that one part of the circuitry couldn't 'talk' to another part, and the noise had reduced greatly. There now remained one flicker line running through the display. Where was it coming from? I was convinced that the power rails were all clean and I had checked for noisy tracks running under the critical circuitry. What else was left to do?

This was quite a puzzler. The only thing left was to eliminate the circuitry until the problem went away. There was a lot of logic circuitry on this board and by removing the clock from part of it, half the circuitry on the board was shut down. The noise was still there. Ok, that meant the problem was not with the logic I had shut down at least. The next worry was the LCD backlight. There was mention in the data sheet that if the frequency was not some prime ratio relative to the horizontal sync period then there could be a display beat. Was that causing the problem? Without the circuit diagram for the backlight inverter, it was difficult to change the frequency in any controlled way, but I did put a pot across one of the resistors and managed to change a frequency on the board by a large amount. No effect, but then there seemed to be two frequencies active on the board.

I would have replaced the inverter with a bench supply, but it was a high voltage AC supply and therefore difficult to replace. The solution I came up with was to remove the backlight from the display completely. Now the display was not visible. I ended up getting a torch and shinning it through a small area of the display. [The torch had to have an incandescent bulb. A fluorescent bench light just wasn't convincing enough!] The noise was still there. There was practically nothing left on the board. Just a single 20 MHz clock to a programmable logic gate on the digital output lines. That couldn't have any effect; it was the wrong frequency and it was nowhere near the critical circuitry.

I shut the 20 MHz clock down and the noise went away. Hurrah! Now that I knew where the noise was coming from it was a simple matter to track down the mechanism. It just so happened that the clock line ran under the phase locked loop capacitors. Now I would have said that this should have caused jitter on the edges of the screen, not brightness bands; I was evidently wrong. In a contest between a theory and the physical universe, the physical facts win every time!

There is a moral to this story. Be analytic to solve the problem at first. Work out what frequency you are hunting for, find the correlated signal and kill it. If this approach doesn't solve the problem, go for the more powerful but simpler elimination scheme. If something is not moving or changing then it is not causing noise. Shut down stages in large chunks until you can make a change to the measured noise. Sometimes it is hard work to shut down a particular stage, but unless you do it, you will have a nagging worry that it is still possibly the source of the problem.

Actually there is an important point that needs to be made about the quality of your investigative procedure; each modification you try must be done carefully. If you incorrectly eliminate a part or area as not being the cause of the noise due to a faulty test (like you decoupled the wrong pin of the IC by mistake), then you can go chasing off into many blind alleys {dead ends}. It is worthwhile to be methodical and careful in your experimental work so that you don't disregard the actual source of the problem. Keep a logbook to record the path you took and what things you have tried. It may be better to repeat a particular experiment at a later stage (and possibly on another day) to prove beyond reasonable doubt that the noise is not coming from that source.

The idea of measuring the thing you are trying to change helps here. If you were monitoring the power rail when you added extra decoupling and nothing happened to the supply rail, you would at least know that something was wrong. If you were merely looking at the noisy output and nothing happened when you increased the decoupling, you might just assume that there was no noise on that particular point.

Now in the measurement method I have said that you should see how much noise you have on any particular rail. Ok, but how much noise is acceptable? If you knew that

50 mV of noise was necessary to cause the amount of effect you were seeing then you could measure the power rail and say it was clean enough, since it was less than 10 mV. If you don't know what the signal sensitivity is at the point you are measuring, then how do you know how low the noise has to be? Obviously you don't.

In the first instance you would just assume that if you could see any noise then that was something to improve. Suppose you had a signal on the final output at 30 kHz, the switched-mode supply frequency. You wonder if the noise might be due to the ripple on the +15 V rail. You measure 26 mV ptp. Is this too much? Who knows? Well just reduce it and see if it has any effect. So you grab the nearest low impedance electrolytic you can lay your hands on and you solder it across the power rails. Don't worry that this case size won't fit, that the component costs too much, and that there is no time to re-lay the PCB. Put those considerations out of your mind. Did the noise on the rail go down? Yes. Good. Did the noise on the output go down? No. Too bad. That wasn't the problem. But *leave the capacitor in place*. When you have finally killed the noise you can pull away the extra components one by one and see if any noise comes back. You may well need several noise fixes on any particular source and if you take off the ones that should work you will not quickly get to the best performance. This is a very important time-saving principle.

The other method is more time consuming. Get a signal generator and deliberately inject a signal into the power supply, or whatever point you are concerned about. Use a coupling capacitor of perhaps 1 nF to 100 nF, depending on the frequency being injected, and the source impedance of the point you are driving. You don't need to inject more than a few tens of millivolts. If this noise appears at the overall output you will be able to calculate the sensitivity of this point in the circuit at the frequency under consideration. Now you will know if your previous measurement technique was sensitive enough to measure the noise that is ordinarily present at that point in the circuit.

Understand that there will never be enough time to entirely eliminate all noise. All you can do is reduce the noise to an *acceptable* level. Perhaps this is below 0.2 LSB or perhaps it is less than 1 least significant digit. Whatever the measure, you will only be allowed, or should only allow yourself, to reduce the noise to some previously determined level. It is often worthwhile setting a target for the noise and aiming for that, rather than having a moving target. You get the noise to 0.7 LSB and then somebody says that's too big. So you spend a week and you halve it. Then the same person says "can't you get it any smaller?" So after a further months work you have halved it again. This can drag on for an undefined time. If you don't have an initial goal, and an acceptable limit, then the work will continue endlessly and you may be accused of failing to meet your timescales. It may be that there are two limits; it must be lower than A, but would be better if lower than $A/3$ for example.

Coupling Mechanisms Summary

- ☹ Finite PSRR of an IC
- ☹ Coupling from a noisy power line.
- ☹ Capacitive coupling from a nearby noisy track.
- ☹ Capacitive coupling from a source of large voltage swings.
- ☹ Mutual inductive coupling from a source of large current swings.
- ☹ Resistive volt drop in a common conductor due to a changing heavy load current.
- ☹ Inductive volt drop in a common conductor due to a rapidly changing load current.
- ☹ Electromagnetic pickup of far-field radiation from external radio sources.
- ☹ Finite surface transfer impedance in a coax cable causing a common-mode current in the screen to be converted to a differential-mode voltage.
- ☹ Imbalance in a cable causing a differential-mode voltage to be partially converted to a common-mode voltage (*longitudinal conversion loss*).
- ☹ Common-mode current in the ground plane.
- ☹ Vibration causing movement of poor joints or generating voltages in *microphonic* components.

Standard Fix Summary

This is the standard list of "fixes" for noise problems. You obviously keep these in your mental 'toolbox', ready to apply to any noise problem that comes up, but there is more to it than that. You can also apply or allow for the fixes *before* you get into trouble. Why do a PCB layout and then find out there are problems when it is being tested? It is sensible to identify possible problems before they occur and to allow contingencies to handle these problems.

Let's take the case of an LCD display backlight. These can have 50 kHz signals at 1 kV levels. This amplitude of signal is a villain looking for a victim. I am not saying that you necessarily need to screen the cables, or screen the inverter, or screen nearby circuitry. These all add cost, but it doesn't cost much to realise that there is likely to be a problem and to include holes on a PCB for a screened lid if one is later found to be needed. Just think ahead as to what might be needed and how the problem could be solved. If you don't do this then the rework cost can be considerable and modelling the next iteration of board will be more difficult.

No *real* problem: be certain that you actually have a problem before you start diving into the circuit. If you have the screens or covers off the amplifier, then you could be picking up signals from nearby unintentional signal sources. These include, but are not limited to, signal generators; computers; CRT displays; fluorescent lights; other test equipment; wireless LAN, keyboard or mouse; and other engineers on nearby benches. You can chase around madly trying to hunt down this "noise source" only to find out that the screened cable from the unused VHF signal generator is close to your circuit. [Screened cables are often not adequate for the purpose of keeping a VHF signal completely contained.] You will save considerable time and mental energy by checking for silly problems like this before attacking your circuits.

Fast moving &/or high frequency voltages: includes line-output transformers (for CRT displays), switched-mode supplies, logic lines, video displays, multiplexed LED displays. Standard fixes are:

- ✓ Brass, copper or tin-plate shield around the source.
- ✓ Brass, copper or tin-plate shield around the sensitive circuit.
- ✓ Keep the fast moving tracks away from sensitive tracks and sensitive areas.
- ✓ Check for optimum ribbon cable routing. Does moving the cable or touching it reduce the noise?
- ✓ Put grounded guard tracks or copper planes between the sensitive circuitry and the noisy circuitry.
- ✓ Use screened wires to carry the fast moving signal.
- ✓ Use screened wires on the sensitive circuitry.
- ✓ Separate the power supply current paths to and from the noisy circuitry so that the current surges do no not flow in a common impedance (shared with the sensitive circuitry).
- ✓ Force currents in the noisy area to stay in that area by using plenty of decoupling capacitors locally.
- ✓ Filter the power lines as they come out of the noisy area so that the noise does not get transferred into the rest of the circuitry. This would involve either a resistor or an inductor in series, with a capacitor to ground.
- ✓ Filter the power lines as they go into a sensitive area.
- ✓ Put cuts in the power & ground planes to discourage noise currents from flowing through the tracks/planes in the sensitive area.
- ✓ Re-position all cables, one at a time, to see if any of them change the noise. Sometimes just touching a noisy cable can reduce the noise! Those cables that do cause noise should be investigated to see if the edge speeds can be reduced, if more grounds are needed in the cable, and if the cable can be re-positioned to minimise the noise.
- ✓ Provide low impedance shunt paths for the high speed currents so they don’t travel through the sensitive analog areas.

Switched or **alternating currents**: Magnetic parts such as CRT deflection coils, switched-mode power supply transformers, switched-mode power supply chokes, relay coils, mains transformers; also output power feeds, memory modules, processors, gate arrays.

- ✓ < 3 kHz magnetic field: mu-metal, iron or tin-plate shield
- ✓ < 30 kHz magnetic field: radio metal, iron or tin-plate shield
- ✓ > 40 kHz magnetic field: copper, aluminium or brass shield, heavily overlapped to give low impedance joints.
- ✓ Electric field of any frequency: copper, brass, aluminium, sprayed metal, carbon based coating (aquadag), conductive cloth, or any other sort of conductor .
- ✓ Twisted pairs (or screened twisted pairs) at source of trouble.
- ✓ Twisted pairs (or screened twisted pairs) at sensitive circuitry.
- ✓ Coaxial cables at source or victim circuitry.

- ✓ Coaxial cable (at source or victim) wrapped around highly permeable core {eg laminated supermetal} for problems below 300 Hz.
- ✓ Small transmit-loop area made by careful track routing of major current paths.
- ✓ Small receive-loop area made by careful track routing of sensitive paths.
- ✓ Re-orientation of wound (magnetic) parts (different axis for pickup or transmission).
- ✓ Toroidal cores for any magnetic components rather than an open magnetic path.

Fast moving currents and voltages generate both electric and magnetic interference and you may need to combine methods to minimise the pickup. For example a screened twisted pair may be needed to eliminate noise from a switched-mode supply.

Axial inductors in filter circuits are excellent inadvertent antennas for magnetic fields. Try changing the direction of the component's axis to see which gives the lowest pickup. You can use the same sort of inductor connected to a length of coax cable to quickly investigate the interfering field and to find a suitable axis for minimum pickup.

Ground Plane Noise

Noise in the ground plane is the hardest type of noise to detect and to correct. It is generally too small to measure directly since we are talking about the volt drop in a continuous sheet of copper. It cannot be 'decoupled' because it is the plane to which you would decouple. You cannot easily model changes because the ground plane is usually on an inner PCB layer and is therefore inaccessible. You have great difficulty establishing that this is really the fault and that only re-laying the board will fix it. If there is that much volt drop along the ground plane then the common-mode noise will be large in comparison. Therefore, even if external measurement equipment would be sensitive enough to measure the noise, the readings will be swamped by the common-mode noise.

The problem is this: current is flowing through the ground plane and is creating different potentials at different points in the plane. In order for this to happen the current must be either very large or at some high frequency. Suppose you have a problem where there is 1 mV being developed across a 2 cm length of ground plane. Let's say the ground plane is 5 cm wide in standard 1 oz copper. The DC resistance of this piece of the plane is $R = 17\times10^{-9}\cdot\dfrac{0.02}{0.05\times34.3\times10^{-6}} = 0.2\,\text{m}\Omega$. To get a 1 mV drop you would therefore need 5 A. Whilst this is possible, it is (hopefully) unlikely that you would have designed the system so badly that you had changes of 5 A occurring in the ground plane. The figure of interest is therefore not the DC resistance but the inductance. The inductance of this piece of the ground plane can be estimated as:

$$L = 2\times\left[\frac{1}{2}+\ln\left(\frac{1.998\times2}{5\times34.3\times10^{-4}}\right)\right]\times2\,\text{nH} = 24\,\text{nH}$$

At 10 kHz the 1 mV drop can be created by 0.7 A. At 1 MHz, only 7 mA is required; this is the problem. Hardly any current makes a real mess of the concept that the ground plane is all at the same potential.

The impedance of a ground plane is very low. It is therefore remarkably difficult to provide an effective shunt path to divert any noise currents. Every effort should therefore be made to return transient currents locally to their source. Shunting/by-passing the sensitive area can then be used for perhaps another 10 dB to 20 dB improvement.

Even if a separate wire is routed alongside the ground plane to avoid the HF current in the ground plane, the mutual inductance coupling between the ground plane and the wire may well be so great that the wire gets an induced volt drop equal to that occurring in the ground plane!

17.9 Pulse Response

If a circuit works well in the frequency domain, there is no guarantee that its performance in the time domain is comparable. It has been established[1] that if the response in the frequency domain peaks to a normalised value of M_P, then the response in the time domain cannot peak by more than 1.18× M_P. This means that if a system is flat in the frequency domain (monotonic roll off), the step-response overshoot cannot be guaranteed to be less than 18%. This is very poor performance for any sort of amplifier!

An ideal ten-pole Butterworth filter is guaranteed to be optimally flat in the frequency domain, and peaks close to the 18% limit in the time domain,. The reason for this can be seen by looking at an amplifier made from 5 cascaded two-pole stages. In order to form a Butterworth characteristic, two of the stages are required to peak in the frequency domain; one of these peaks by a factor of 3×. With one stage so heavily peaked, it should not be surprising that the overall output overshoots in the time domain.

One might suppose that a system which had a flat pulse response would then be equally flat in the frequency domain; unfortunately this is not true either! Consider the square-wave response {step response} of a system. The square-wave can be decomposed into its constituent Fourier components, in other words the square wave can be considered to be composed of an infinite sum of harmonically related sinewaves of ever-decreasing amplitude.

$$\boxed{\text{square wave} = 0.5 + \frac{2}{\pi} \cdot \sum_{n=1,\,3,\ldots}^{\infty} \frac{1}{n} \cdot \sin(n\omega t)}$$

This is the Fourier series for a square wave going from 0 to 1 unit. Text books usually give the harmonic coefficients as $1/(2n+1)$ and use steps of 1, thereby complicating the equation. Using a step size of 2 makes the equation simpler and the coefficients easier to understand.

Suppose the frequency response has a narrow 4% peak at 50 kHz. A 10 kHz square wave will have its fifth harmonic peaked by 4%, but the fifth harmonic is only 1/5th the amplitude of the fundamental; the time domain peaking can be more like 0.5%. If the square fundamental is such that no harmonic hits the resonant frequency, then *the frequency domain peaking will be undetectable on that pulse response test*. It is therefore important to check broadband systems at multiple frequencies to check for lumps and bumps in the frequency flatness. If the system is to be used in the time domain, it is additionally important to check at several repetition frequencies of the square wave.

Some texts describe ‘distortion’ of a time domain signal [pulse] as any change from its original state. By this definition every amplifier, every cable and every signal path introduces distortion onto a signal. This is not a useful definition. I am going to define ***aberration*** on a step response as any *unusual* feature added to the waveform. To be more

[1] A. Papoulis, '5-3 Evaluation of the Step Response.' in *The Fourier Integral and Its Application* (McGraw-Hill, 1962), pp. 89-93.

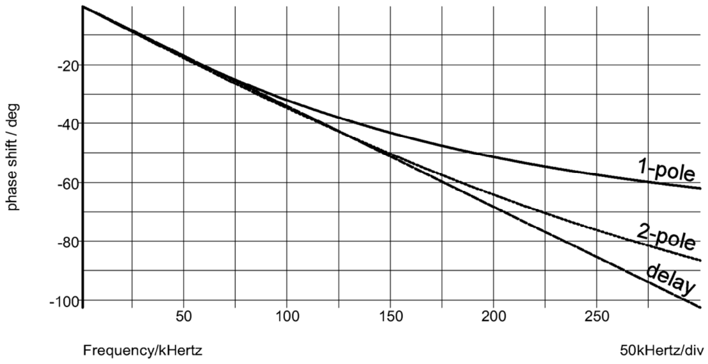
explicit, any bumps, wiggles, droops, kinks, holes; whatever you want to call them, anything added to the waveform that makes it look non-ideal. Get the idea that a rising edge that is merely slowed down has not been aberrated. This is a perfectly usual and expected response when an additional stage has been placed in the circuit.

There is a widespread opinion in this field that *phase linearity* is the key goal of an amplifier or system in order to achieve an excellent pulse response. Notice that I am using a term from the frequency domain to qualify the performance in the time domain.

The idea that phase linearity is the ideal to aim for, comes from the fact that the only network which has no adverse effect on a step response is a pure theoretical time delay circuit such as an ideal lossless coaxial cable. The phase shift increases linearly with frequency as you saw in the section on oscillators. $\left|\frac{d\phi}{df}\right| = 2\pi\tau$. If the phase shifts linearly with frequency then you have a pure time delay element; there is neither distortion nor aberration of the signal. Great, but such a system does not exist and this 'measure' does not say how close to ideal phase linearity you have to get in order to achieve a certain degree of aberration.

A single-pole filter does not aberrate a pulse response. In fact it gives a very clean pulse response. If you wanted a measure of a good pulse response then a single-pole filter would fit well. A two-pole *synchronously tuned* filter (both poles equal) also gives a very clean pulse response, but what are the phase responses like?

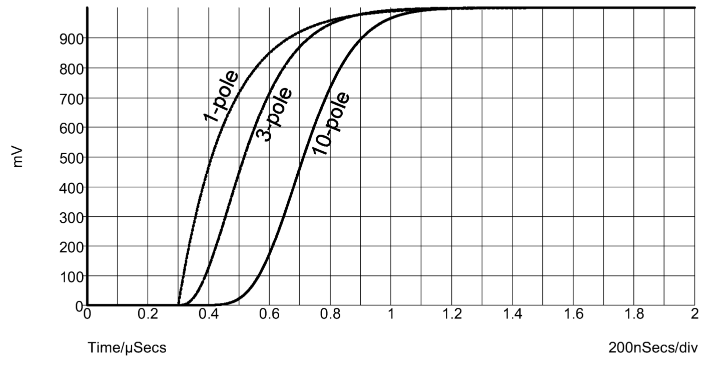
FIGURE 17.9A:



It is evident that the phase response of the filter gets closer to the simple linear phase response of an ideal delay line as the number of poles increases. Notice that these plots have been normalised to the phase performance and not to the bandwidth.

If the pulse responses of these networks are compared, it is seen that the initial part of the rising edge gets progressively more rounded, whilst the final part of the rising edge gets progressively less rounded as the number of poles increases. The centre of the rising edge is also more delayed as the number of poles increases.

FIGURE 17.9B:



These step responses are all normalised for 1 MHz bandwidth. They are for one pole, three pole and ten pole networks of synchronously-tuned, cascaded, buffered, single-pole sections. The simulated risetimes change smoothly from 350 ns for the single-pole system to 340 ns for the ten-pole system.

It is found in practice that the risetime-bandwidth product is relatively constant at ≈350 ns×MHz for all networks having good pulse response characteristic. This product is usually expressed as the dimensionless constant 0.35; it is used to estimate the risetime from the bandwidth, and vice-versa.

The simplest mathematical form of an ideal shape for a pulse response is a ramp from the minimum to the maximum values, often referred to as a trapezoidal waveform. The question then arises as to how to quantify the deviation from this ideal, as might be required for computer optimisation for example. In control system engineering some such indices are already used. If the error is denoted by an error function, *e(t)*, these performance indices include:

Integral of the **A**bsolute value of the **E**rror,

$$IAE = \int_0^\infty \left| e(t) \right| \cdot dt$$

Integral of the **S**quare of the **E**rror,

$$ISE = \int_0^\infty e^2(t) \cdot dt$$

Integral of (**T**ime × **A**bsolute **E**rror),

$$ITAE = \int_0^\infty t \cdot \left| e(t) \right| \cdot dt$$

Integral of (**T**ime × **S**quared **E**rror),

$$ITSE = \int_0^\infty t \cdot e^2(t) \cdot dt$$

Integral of (**T**ime **S**quared × **A**bsolute **E**rror),

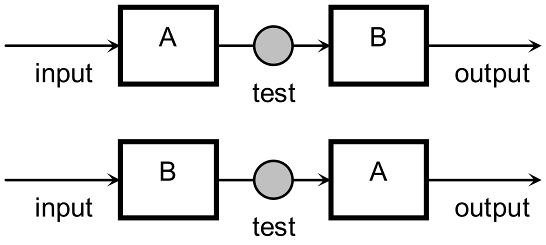
$$ISTAE = \int_0^\infty t^2 \cdot \left| e(t) \right| \cdot dt$$

Integral of (**S**quared **T**ime × **S**quared **E**rror),

$$ISTSE = \int_0^\infty t^2 \cdot e^2(t) \cdot dt$$

It is not possible to say that one measure is 'better' than any other, they are just different and each has its own specific optimum application circumstance.

FIGURE 17.9C:



If you have two cascaded amplifiers, each having input impedance Z_{IN}, and output impedance Z_{OUT}, the pulse response out of the final output will be the same, regardless of the order in which the two stages are connected.

This is obvious if you think about the system in the frequency domain and just multiply the transfer functions. For time domain blocks it is much more difficult to think about the effect of combining pulse responses, which is the point of this discussion. If the frequency domain responses are the same in both amplitude and phase, the time domain responses must also be the same.

Let's suppose that you can successfully probe the intermediate test point with a high impedance probe that does not appreciably load the circuit. You might be very upset with the design if you saw 30% overshoot at the intermediate point for an overall flat pulse response. However, by swapping the order of the sections the overshoot miraculously disappears. This has not made the system any better, however. The overshooting stage is still present, *you just don't notice it* because you are no longer driving it with a fast enough edge. A system which has 30% overshoot on a stage may not be as stable with time or between units as you would like, but this also applies to the case where the overshoot is not readily apparent.

The best way to spot this condition is to test individual stages with a fast pulse. Another thing you can do is to measure the pulse risetime going in and coming out. If the risetime is faster coming out than going in, you know a stage is heavily peaked. This is not necessarily a problem. Just be aware of what your circuit is doing and don't be fooled by what appears to be clean intermediate responses.

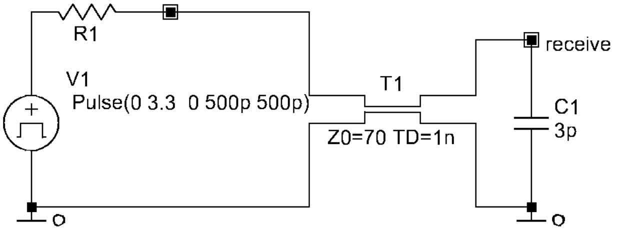
I have previously mentioned that the return-to-zero edge of a pulse is always the best from a standardisation point of view, the high level tending to have some sort of droop or time-constant associated with it. When current is drawn from a power rail the output impedance will give a small, but possibly significant, volt drop. Having reduced this source impedance as much as possible, and added as much decoupling as you can get, the rail may still not be of sufficiently low impedance. The solution is compensation. By applying an equal, but opposite, current pulse to the power rail, the power rail's source impedance becomes less critical. The current pulses will never be totally opposite in amplitude or phase/position, but the resulting power rail transient should be improved by at least a factor of 10×.

17.10 Transmission Lines

Fast digital systems have to be considered as analog systems not digital systems. Actually 'fast' really means that the risetime of the signal is faster than the transmission line propagation delay. The speed of light, and other electromagnetic radiations, in free space is 3×10^8 m/s. Inverting this gives a transit time of 3.3 ns/m. When electrical signals travel down cables and PCBs they travel slower than this maximum because of the dielectric constant of the material. A good estimate is 5 ns/m, in the absence of more accurate data.

There are many rules of thumb for deciding when the line is sufficiently long to worry about the reflected signals. First let's model a single gate driving a single high impedance input.

FIGURE 17.10A:



V1 and R1 model the sending-end gate. C1 models the receiving-end gate. 1 ns delay models 20 cm of PCB track. The characteristic impedance of the PCB track has been put at 70 Ω. This will range around 30 Ω – 200 Ω, according to the track width and the spacing to the ground plane or nearby ground tracks.

FIGURE 17.10B:

It should be obvious that the resulting pulse response shown to the right gives a completely unworkable digital system as one rising edge could actually generate three rising edge transitions.

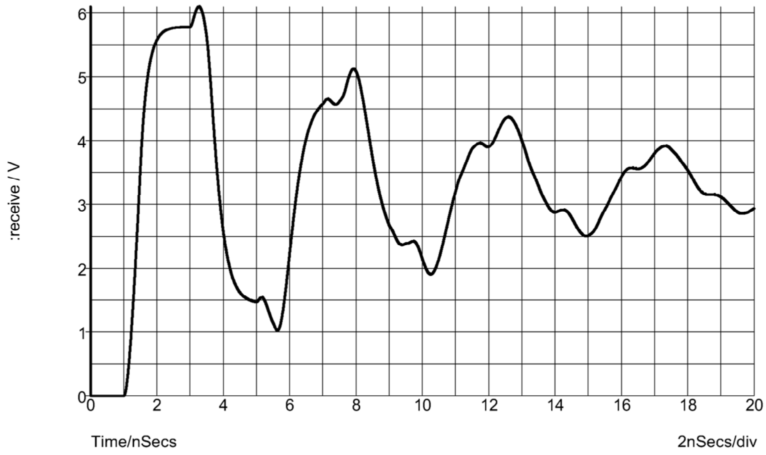
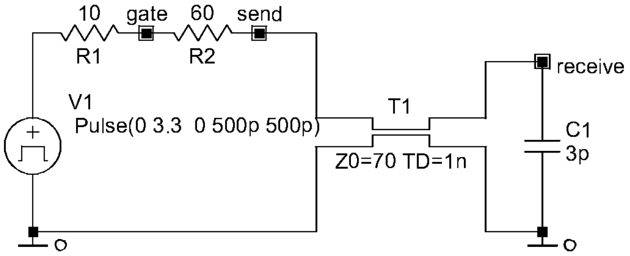


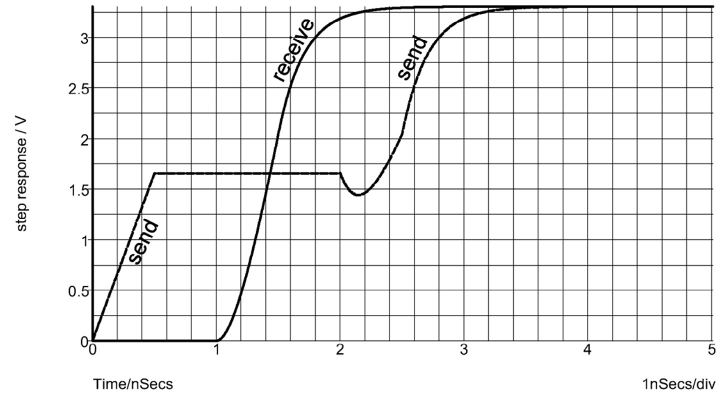
FIGURE 17.10C:



The addition of a resistor at the sending end (R2) is all that is necessary to fix this system. The sum of the added resistor and the internal gate source resistance are made to match the impedance of the transmission line, usually by direct experiment.

FIGURE 17.10D:

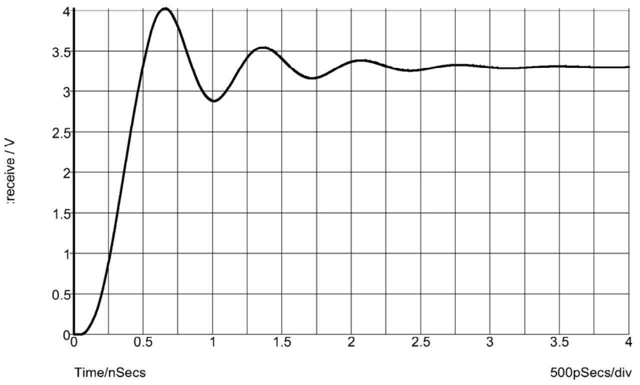
The receiving end signal is now very clean. The sending end signal, however, is completely unacceptable for any gate *input*. The characteristic transient half-amplitude level lasts for double the delay time of the transmission line.



When you first apply a step edge to a transmission line, a current flows. The current is the voltage divided by the *surge impedance* (characteristic impedance) of the line. Therefore, for a correct series termination, the voltage seen at 'send' is initially half the step amplitude because of the resistive divider effect of R1 + R2 and the surge impedance of the line.

When this current surge gets to the end of the line, it encounters a load. If the load is matched to the line, ie it is a resistor equal to the characteristic impedance of the line, the current pulse is absorbed. If the load impedance is higher than the characteristic impedance then you can view this as an excess of current. For example, if the line is open-circuit then no current can flow. This non-flowing of current is then communicated to the sending end of the line by sending a negative current reflection back down the line. Putting a suitable resistor in series with the line at the sending end is known as *series termination* and the qualitative criterion for not using the resistor is that the transmission line is "short".

FIGURE 17.10E:



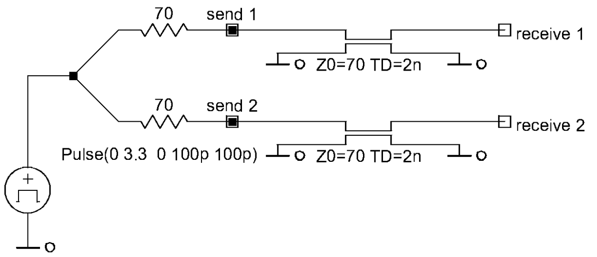
In this case the transmission line has been reduced to 50 ps, but without the extra series termination resistor. The transmission line is one tenth the length of the rising edge and yet there is still a substantial ring! If the line were any longer, the first dip after the overshoot might re-cross the logic threshold, causing a spurious event.

The capacitive load on the end of the line is making the situation considerably worse. If the capacitive load is ten times lower, the ringing is less bad so an edge-speed to delay ratio as low as 4 could be acceptable. As a guide, about 0.5 pF of load capacitance per ns of risetime is tolerable before the edge-speed/delay ratio needs to go from 4 to 10. In other words for a 2 ns edge speed, 1 pF of load capacitance is ok if the trace is 500 ps long. In practice this means that most lines faster than 2 ns will have to be carefully checked for length and terminated when necessary.

For any digital line longer than one tenth the edge speed, consider using a series or parallel termination scheme.

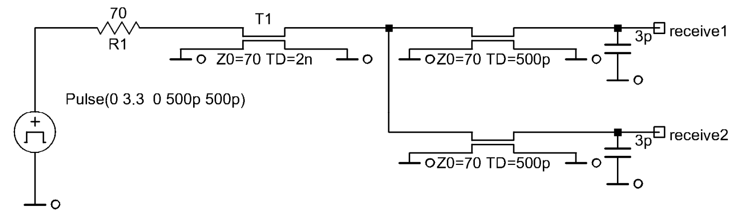
The same rules apply to 5 µs signals as apply to 500 ps signals; it is just that the line lengths that cause trouble are 10,000× longer. The capacitive loading effect is unlikely to cause a problem because a load of 2.5 nF would have to be deliberate.

FIGURE 17.10F:



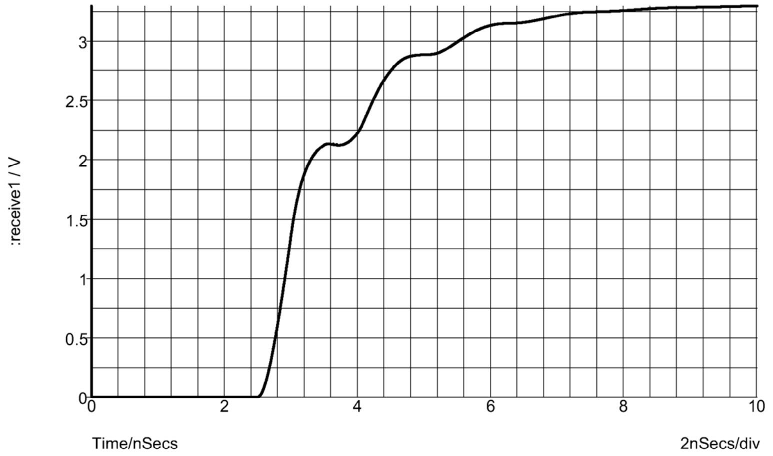
When you have to drive multiple gates from one source, the solution is either to parallel terminate, or to use separate series termination resistors for each of the receiving gates. The only exception to this rule is if the gates are close together compared to the signal risetime. In this case it is possible to split the line and feed both gates. You should ideally make the path length equal to each of the gates from the split. This will minimise pulse irregularities and will also minimise ***skew*** between the edges as well. The line will not be correctly matched, but the difference may be small enough that it doesn't cause problems.

FIGURE 17.10G:



This simulation shows the limit of acceptability. Each stub has the same delay as the edge speed. The length of the main line is irrelevant because it is correctly matched to the sending end impedance. A return pulse travelling right to left in T1 will be absorbed completely in the series termination R1.

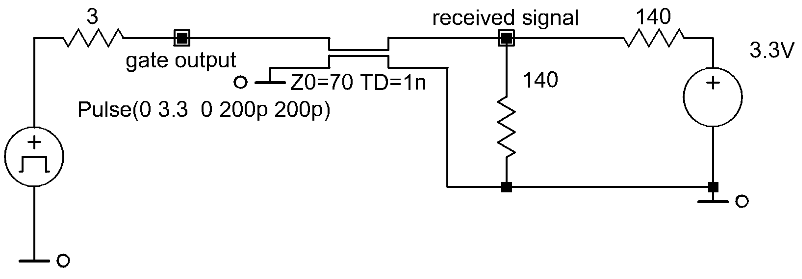
FIGURE 17.10H:



Notice that the rising edge is bordering on ***non-monotonic*** behaviour close to the switching threshold. You get a perfectly matched response if the characteristic impedance of the stubs is exactly double that of the main line, but this may not be achievable in practice.

The straight forward way of applying a parallel termination is to put a resistor to ground at the receiving end. The resistor is equal to the characteristic impedance of the line and will be in the range 30 Ω – 200 Ω. Parallel termination is therefore more difficult than series termination in terms of loading. Putting a 70 Ω DC load on a CMOS gate will almost certainly be more than it can drive. An ECL gate, on the other hand, would not have a problem with this [provided that the termination is taken to the –2 V rail]. If this were 3.3 V logic then the high level drive current for the parallel termination would be 45 mA.

FIGURE 17.10I:



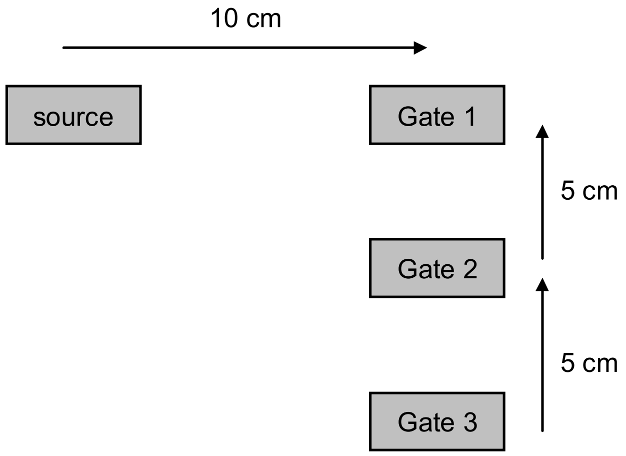
The high level drive current can be halved by using a so called *Thévenin termination* as shown here. (The logic gate is still connected to the "received signal" node.)

The Thévenin equivalent load impedance correctly matches the line impedance. Although the load current on the sending-end gate is halved, the sending-end gate now has to be able to both source and sink equal amounts of current.

For analog systems it is possible to use either termination method and sometimes the line is *double terminated* to get the best performance. This means that it is matched at both ends of the line. Double termination is not possible with single-ended logic families, however, because the signal level will be halved. Differential logic families such as ECL , PECL and LVDS, can be double terminated, providing they use a differential line receiver [or a gate with a differential input] to regenerate the pulse.

***EX 17.10.1**: The digital guys need a hand. They have a 200 ps edge-speed clock to transmit to three gates spaced out in different areas of a PCB. The layout is roughly as shown below.

FIGURE 17.10L:



They want to wire this circuit up to get good pulse edges and to minimise skew. One suggestion is to run the feed from the source directly to gate 1, then follow on down to gates 2 and 3, with a parallel termination at gate 3. Another suggestion is to series-terminate at the source and feed to gate 2, with stubs going off to gates 1 and 3 of equal length.

What is your advice?

When using coaxial interconnects between different PCBs it is possible to play games with the series and parallel termination values to optimise the pulse response. This tuning is not restricted to just changing the resistor values up and down. It is possible to add small capacitors and/or inductors into the circuit to minimise overshoot, peak up an undershoot, and generally fine-tune the pulse response and/or frequency response.

17.11 Filters

The subject of filters is more than enough to completely fill a large book. If you have to do a lot of work with filters having more than three poles, you definitely need to get one or more books on the subject and/or some filter design software. However, the first thing you need to establish is your need for more information on the subject, or indeed your lack of need for more information than is contained here.

***EX 17.11.1**: What is the purpose of an electrical/electronic filter?

The whole idea of filtering has to do with the terms *signal* and *noise*. The signal is what you want; the noise could be said to be anything else that you don't want. In a telephone system where individual telephone messages occupy their own band of frequencies, it is clear that you want to separate out one particular telephone conversation from the "noise" of all the other conversations.

An ideal filter would therefore pass the signal without attenuating or distorting it at all; this filter would also give sufficient attenuation to the noise to make it imperceptible. Now this is a tough spec and in practice it is impossible to meet. A filter will distort the signal at least a small amount, and noise will get through to at least a small extent as well. Obviously the more components you use, and the more design time you expend, the better the filter can perform to the original spec. It is the same old fundamental of design; just how much is somebody prepared to pay to get a given performance?

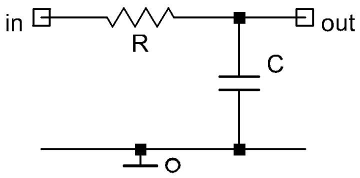
There are five basic filter types, classified according to their *frequency domain* characteristics:

- Low-pass (attenuates higher frequencies; passes DC and low frequencies)
- High-pass (attenuates DC and lower frequencies; passes high frequencies)
- Band-pass (called notch-pass filters when the band is very narrow)
- Band-stop {band-suppression} (narrow band type called notch-filter)
- All-pass equalisers (produce phase shift rather than amplitude change)

You will notice that even filters used for shaping pulses (time-domain) are actually specified in the frequency domain.

In handbooks on filters it is usual to give low-pass 'prototypes'. This is a low-pass filter which has been designed using a source resistor of 1 Ω and a corner frequency of 1 rad/s. The values are then scaled according to the actual frequency and impedance in the final circuit. The low-pass prototype can also be transformed to a high-pass or band-pass configuration by the rules given.

FIGURE 17.11A:



This is the simplest low-pass filter for use from a low source impedance when feeding into a relatively high impedance input. It passes DC and attenuates frequencies above the *corner frequency*. Swapping the R and C over gives a high-pass filter. Filter handbooks typically use an inductor in place of the resistor since they show circuits designed to drive low input impedance (typically 50 Ω, 75 Ω or 600 Ω) stages. In filter handbooks you are given the low-pass form and you have to work out the high-pass and band-pass forms from that circuit.

EX 17.11.2: Assume that the above filter is driven from a source impedance of zero and drives a load impedance of infinity. Take the 3 dB bandwidth of the filter as B and use f as the frequency of operation.

a) Write an equation for the magnitude transfer response of the filter.
b) What is the magnitude response at a frequency equal to half the bandwidth?
c) What is the magnitude response at a frequency equal to double the bandwidth?
d) What is the theoretical magnitude response at a frequency equal to 10,000× the bandwidth?
e) Why was the word 'theoretical' added in the previous question?

The single-pole filter attenuates the signal well before the corner frequency; if you need <0.1% loss, the signal frequency has to be at least a factor of 22.4× lower than the 3 dB bandwidth of the filter. This filter corner is often described as 'soggy', soft, or not 'sharp'. To get a sharper corner you need more poles, a higher *order* filter.

The simplest concept is to cascade several of these single-pole stages. The stages need to be buffered from each other, however, as each stage would otherwise significantly load the previous stage. This type is of filter is known as a *synchronously tuned* filter; all the poles are real {no complex conjugates} and equal.

The magnitude of the transfer response of an N-pole synchronously tuned filter is not very nice to calculate:

$$|T| = \frac{1}{\left[\sqrt{1+\left(\frac{f}{B'}\right)^2}\right]^N}$$

The problem is that the bandwidth B′ in the formula is the 3 dB bandwidth of the individual stages and not of the overall filter. You know that at the band edge of the filter, the magnitude response is $\frac{1}{\sqrt{2}}$. Thus:

$$\left[\sqrt{1+\left(\frac{B}{B'}\right)^2}\right]^N = \sqrt{2}\,,$$ giving the rather unpleasant equation $B = B' \cdot \sqrt{\sqrt[N]{2}-1}$.

The notation used is $\sqrt[N]{2} \equiv 2^{1/N}$. Consider a 4-pole synchronously tuned filter at frequencies of B/2 and 2B.

At B/2, $|T| = \frac{1}{\left[\sqrt{1+\frac{\sqrt[4]{2}-1}{4}}\right]^4} = 0.9117$ At 2B, $|T| = \frac{1}{\left[\sqrt{1+4\cdot\left(\sqrt[4]{2}-1\right)}\right]^4} = 0.3240$

The result is better than that achieved with the single-pole filter, but not by a great deal. This is the whole point of filter theory; to obtain a better response than that given by simple cascaded stages.

The first filter to look at is the ***Butterworth*** filter.[2] Really it is whole class of filters, because the Butterworth principle applies to filters of any arbitrary order. The idea is extremely simple; you make the magnitude of the transfer function:

$$|T| = \frac{1}{\sqrt{1+\left(\frac{f}{B}\right)^{2N}}}$$, where N is the order of the filter {the number of poles}.

This is a very easy function to work with and, like the synchronously tuned filter, the amplitude reduces monotonically with frequency. This table of transfer response magnitudes shows how useful the Butterworth is compared to the two examples previously given.

Filter type	Gain at B/2	Gain at 2B
Single-Pole	0.8944	0.4472
Four-Pole Synchronous	0.9117	0.3240
Two-pole Butterworth	0.9701	0.2425
Three-pole Butterworth	0.9923	0.1240

The Butterworth response is remarkably sharp, and considerably better than any synchronously tuned filter section. There are two penalties though; firstly the circuit is more difficult to design (from first principles) and secondly, the phase response is not ideal, resulting in overshoot and ringing on the step response.

The first point was that the circuit is difficult to design from first principles. The answer is not to design the filter from first principles! Get out a standard filter table handbook, or a computer program, and use the values given.

The second point is more difficult. The step response of a Butterworth filter overshoots by 8% for a 3-pole filter, rising steadily to 17% for a 10-pole system. This poor pulse response often makes the Butterworth filter unsuitable for time-domain

[2] S. Butterworth, 'On the Theory of Filter Amplifiers', in *Experimental Wireless & The Wireless Engineer* (Oct 1930), pp. 536-541.

systems. If a non-overshooting response is required then either a ***Gaussian*** or a ***Bessel-Thomson*** filter is the answer. These give better time domain responses, at the expense of slower roll-offs in the frequency domain. The Gaussian gives virtually no overshoot, but has a slower roll-off than the Bessel.

The Butterworth attenuation characteristic is so simple, mathematically, that it is very quick to work out what order of filter is required at any particular frequency. For three pole filters and above, when $f \geq 2 \cdot B$ you get a simple formula for the attenuation characteristic:

$$|T| = \frac{1}{\sqrt{1+\left(\frac{f}{B}\right)^{2n}}} = \left(\frac{B}{f}\right)^n \cdot \frac{1}{\sqrt{1+\left(\frac{B}{f}\right)^{2n}}} \approx \left(\frac{B}{f}\right)^n \cdot \left(1-\frac{1}{2}\left(\frac{B}{f}\right)^{2n}\right) \approx \left(\frac{B}{f}\right)^n$$

Under these condition the formula is accurate to better than 0.8%.
This gives a simple formula for the attenuation in dB under these same conditions:

$$\boxed{\text{Butterworth Attenuation (dB)} = 20 \cdot n \cdot \log_{10}\left(\frac{f}{B}\right)}$$

If you need 60 dB attenuation at 10× the bandwidth then you need a 3-pole Butterworth. It doesn't get any easier than that!

Now there are a lot more complexities that need to be added into this picture. If you are designing a filter for insertion into a 50 Ω transmission system then the load and source impedances are going to be 50 Ω. If, on the other hand, the filter is for use inside a piece of equipment, the output of the filter might be buffered. This apparently subtle difference changes the values that would be used in the filter circuit completely. In fact the idea that the source and load impedances could be different itself gives an infinite number of possibilities.

The factors that have to be considered are:

- ✓ The source impedance.
- ✓ The load impedance.
- ✓ The order of the filter (number of poles).
- ✓ The type of filter characteristic (synchronously tuned, Butterworth, Bessel, Gaussian, Chebyshev &c).
- ✓ The insertion loss of the filter, both in-band and out of band
- ✓ The reflection coefficient of the filter, both in-band and out of band.

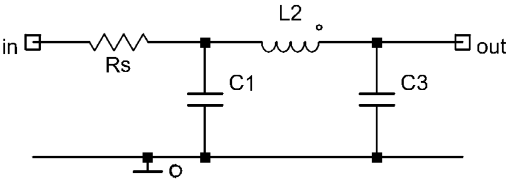
Perhaps now you can understand why whole books of tables of filters are required to fully handle this situation. I am going to stay clear of this complexity and give you a few simple solutions to simple problems.

The Chebyshev filter is going to be given in the following sections. It gives a more aggressive rolloff than the Butterworth, the penalties being a series of ripples in the passband and more pulse ***aberration***. The "0.5 dB Chebyshev" chosen has 0.5 dB peak-to-peak ripples in the passband, hence its name.

Let's look at the case of the unloaded filter; by this I mean that there is a buffer following it so that the load is not very heavy. Indeed, if the input impedance of the buffer is predominately capacitive, this capacitance can be used as part of the final capacitive element in the filter. For the following networks of inductors and capacitors, the *order* of the filter is simply the number of reactive elements.

3-pole low-pass filters:

FIGURE 17.11B:



The table gives values for a corner frequency of 1 rad/s and a source resistance of 1 Ω. This is the way they are presented in many filter tables.

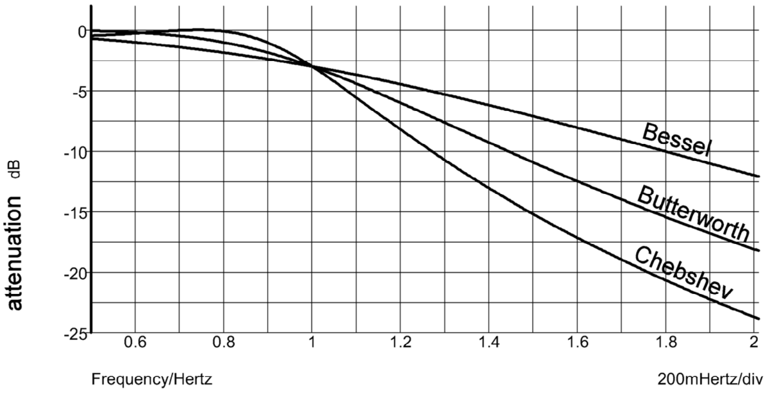
characteristic	C_1'	L_2'	C_3'	Attenuation at $2\cdot B$	Rising Edge Step Response
Bessel	0.2926	0.8427	1.463	12.0 dB	+0.7%; –0.0%
Butterworth	0.5000	1.333	1.500	18.1 dB	+8.2% ; –1.5%
0.5dB Chebyshev	0.9318	1.518	1.572	23.7 dB	+8.9% ; –6.8%

$$C_1 = \frac{C_1'}{2\pi B \cdot R_S}\ ; \qquad L_2 = \frac{R_S L_2'}{2\pi B}\ ; \qquad C_3 = \frac{C_3'}{2\pi B \cdot R_S}$$

The step response figures show the overshoot as the positive percent value, with the dip of the subsequent ring (hook) shown by the negative percent figure.

The Bessel has a very soft corner compared to the Butterworth and Chebyshev responses. After the *2·B* frequency point, all three responses have a slope of approximately 60 dB/decade (the asymptotic slope of a 3-pole filter).

FIGURE 17.11C:



This third order Chebyshev low-pass filter has one 0.5 dB trough, relative to the DC level. The rising back to the 0 dB level is considered a peak when counting peaks + troughs.

In general, a Chebyshev filter of order *n* has *n*–1 peaks + troughs. Also note that these are general curves for third-order filters. The impedances at input and output affect the components used, but do not change the response curves shown above.

EX 17.11.3:

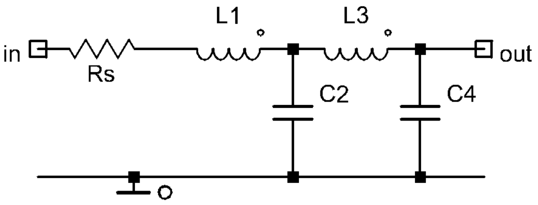
a) Using the design table above, calculate the values for a low-pass 3-pole Butterworth filter to run from a source impedance of 10 kΩ, with a corner frequency of 10 kHz.

b) Using the design table above, calculate the values for a low-pass 3-pole Butterworth filter to run from a source impedance of 100 Ω, with a corner frequency of 10 kHz.

4-pole low-pass filters:

The table gives values for a corner frequency of 1 rad/s and a source resistance of 1 Ω.

FIGURE 17.11D:

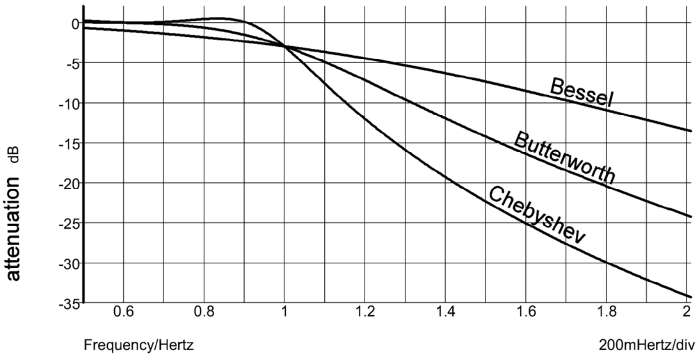


Characteristic	L_1'	C_2'	L_3'	C_4'	Attenuation at $2{\cdot}B$	Rising Edge Step Response
Bessel	0.2114	0.6127	0.781	1.501	13.4 dB	+0.8%; –0.0%
Butterworth	0.3827	1.082	1.577	1.531	24.1 dB	+10.8% ; –3.0%
0.5 dB Chebyshev	0.9239	1.539	1.911	1.453	34.1 dB	+18.1% ; –4.4%

$$L_1 = \frac{R_S L_1'}{2\pi B}; \quad C_2 = \frac{C_2'}{2\pi B \cdot R_S}; \quad L_3 = \frac{R_S L_3'}{2\pi B}; \quad C_4 = \frac{C_4'}{2\pi B \cdot R_S}$$

The step response figures show the overshoot as the positive percent value, with the dip of the subsequent ring (hook) shown by the negative percent figure.

FIGURE 17.11E:



The forth-order Chebyshev has two peaks of 0.5 dB relative to the DC response. Only one of these peaks is visible on the scale shown. The settling back to the 0 dB level in between these two peaks is considered as a trough when counting peaks + troughs.

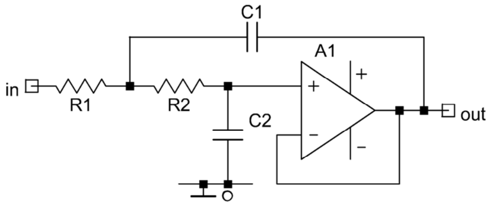
In general, a Chebyshev filter of order n has n–1 peaks + troughs. Also note that the above curves apply in general to fourth-order filters.

The point to be aware of with these filter tables and simulations is that they assume ideal components. If you approach or exceed the self-resonant frequencies of the inductors and capacitors, the filter response must necessarily be very different from the simulated response. In particular, the Q of the inductors and capacitors needs to be higher for the faster roll-off filters. Component Q needs to be better than 3 for a Bessel filter, better than 15 for a Butterworth, better than 39 for a 0.1 dB Chebyshev, and better than 75 for a 1 dB ripple Chebyshev.[3] (The 1 dB Chebyshev has a faster rolloff and therefore needs higher Q components than the 0.1 dB Chebyshev.)

Active Filters:

The passive filters given so far are quite safe in terms of their performance with respect to signals in excess of the passband. Nothing unpleasant happens unless the capacitors and inductors become self-resonant. With active filters, on the other hand, the amplifier characteristics have to be taken into account, not least of which is because slew rate limiting can cause DC offsets.

FIGURE 17.11F:



This is the *Sallen-Key* filter.[4] It is intended to have good DC accuracy because of the opamp. Since it uses no inductors, it is useful for low-frequency operation and with modern opamps this operation can extend up to 10 MHz.

The Sallen-Key stages can be cascaded to get any even number of poles, although all the stages will be different. The maths is complicated, but free software packages are available which present all the component values and transfer curves in a very easy form.[†]

One problem that is seldom mentioned in text books is the out-of-band performance for a large signal. If the filter is intended for 10 kHz operation, the opamp may be a low power, low GBW, low slew-rate device. What happens when the input gets blasted with a full-scale 20 MHz signal? The answer is not very well defined. Slew-rate limiting and rectification in the opamp are possible, since it has to sink the initial surge of current through R1 then C1.

R1 could be made so large that the current in question is quite small. The problem with this is that the impedances of R1 and R2 have quite strict bounds. There is obviously a ratio of the resistors and capacitors to consider in terms of the cut-off frequency of the filter. Then there is the loading effect on the opamp, which means that having R1 and R2 below say 100 Ω can make A1 struggle. The answer is to put R1 and R2 up into the 10 kΩ range, minimising the loading effect on A1. But then you need a FET input opamp. Another undocumented feature also comes into play. The input impedance of a JFET amplifier is not as “infinite” as you might expect.

More particularly the input capacitance, which you may have taken into account by

[3] C. Bowick, 'Filter Design', in *RF Circuit Design* (Sams, 1982), pp. 44-65.

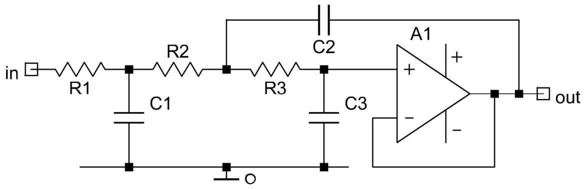
[4] R.P. Sallen, and E.L. Key, 'A Practical Method of Designing RC Active Filters', in *Institute of Radio Engineers: Transactions on Circuit Theory*, CT-2 (March 1955), pp. 74-85.

[†] for example FilterPro from Texas Instruments/Burr-Brown.

lowering C2, is non-linear since it usually includes one or more a reverse biased diode junctions. Many FET input opamps therefore suffer from dominantly second harmonic distortion when used from high impedance sources (>10 kΩ) at quite modest frequencies (>10 kHz). This is hidden in opamp data sheets by using 1 kΩ source impedances when doing THD plots. A possible solution for this problem is to put one or more reverse biased diodes across the input as a compensation network. For a p-channel JFET opamp the diodes only go up to the positive power rail. This trick may result in a 6 dB to 10 dB improvement in second harmonic distortion due to the non-linear input capacitance.

It is better to play safe and to not run devices beyond the point where their performance can comfortably be predicted. This is not always possible, but for the sake of an extra resistor and capacitor I would do it. It is a simple matter to put a passive filter in front of the Sallen-Key filter. Let the passive filter take out some of the fast transient before the active filter gets blasted by it. Now you could put in a small RC filter with a corner frequency more than a decade away from the 2-pole corner. This would be a simple approach. The other approach is to add in these components in such a way as to get a full 3-pole filter.[5] The only difference between these approaches is the values of the resistors and capacitors.

FIGURE 17.11G:



The maths involved in working out the values is not pleasant. One answer is to use a table of values. An alternative approach is to use a mathematical software package such as Mathcad™ to solve the simultaneous equations and get the desired coefficients for the type of filter that you want. The transfer function for this filter works out as:

$$\frac{V_{OUT}}{V_{IN}} = \frac{1}{A_3 \cdot s^3 + A_2 \cdot s^2 + A_1 \cdot s + 1}$$

where:

$$A_3 = \mathrm{C1} \cdot \mathrm{C2} \cdot \mathrm{C3} \cdot \mathrm{R1} \cdot \mathrm{R2} \cdot \mathrm{R3}$$

$$A_2 = \mathrm{C2} \cdot \mathrm{C3} \cdot \mathrm{R2} \cdot \mathrm{R3} + \mathrm{C1} \cdot \mathrm{C3} \cdot \mathrm{R1} \cdot \mathrm{R3} + \mathrm{C2} \cdot \mathrm{C3} \cdot \mathrm{R1} \cdot \mathrm{R3} + \mathrm{C1} \cdot \mathrm{C3} \cdot \mathrm{R1} \cdot \mathrm{R2}$$

$$A_1 = \mathrm{C1} \cdot \mathrm{R1} + \mathrm{C3} \cdot \mathrm{R2} + \mathrm{C3} \cdot \mathrm{R3} + \mathrm{C3} \cdot \mathrm{R1}$$

Some texts suggest that you set all the resistors equal and work out the capacitor values. That is a remarkably stupid approach, since these capacitors will be between 10× and 50× the cost of resistors. They are also difficult to get in anything other than E6 values, whereas resistors can be obtained in E96 values without difficulty. The key thing is to get the capacitors in nice easy to obtain values, then fiddle the resistors to suit. This may require some series or parallel combinations, but it is much cheaper doing this fiddling with resistors rather than capacitors.

For a third-order system, it is convenient to scale the frequency so that the coefficient

[5] P.R. Geffe, 'How to Build High-Quality Filters Out of Low Quality Parts', in *Electronics*, 49, no. 23 (Nov 1976), pp. 111-113.

of s^3 is unity. It is also convenient to consider the non-s term as unity, making the DC gain unity. This simplifies the transfer function of the filter to:

$$\frac{V_{OUT}}{V_{IN}} = \frac{1}{s^3 + A_2 \cdot s^2 + A_1 \cdot s + 1}$$

The coefficients for the various filter types are given here for reference, but the table below gives actual resistor and capacitor values.

Characteristic	A2 (coefficient of s^2)	A1 (coefficient of s)
Bessel	2.433	2.466
Butterworth	2.000	2.000
0.5 dB Chebyshev	1.401	1.918

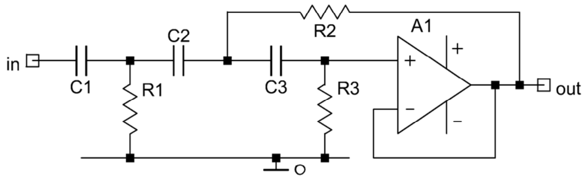
In order to scale the values to a given corner frequency, scale all the resistors by a common factor and/or all the capacitors by a common factor; these two factors need not be equal. To increase the frequency by a factor of 10×, for example, you could reduce all the resistors by a factor of 10×. Or, you could reduce all the capacitors by a factor of 10×.

1 kHz (low-pass)	R1	R2	R3	C1	C2	C3
Bessel	19.13 kΩ	13.84 kΩ	54.66 kΩ	10 nF	10 nF	1 nF
Butterworth	20.56 kΩ	33.32 kΩ	58.86 kΩ	10 nF	10 nF	1 nF
0.5 dB Chebyshev	33.71 kΩ	42.66 kΩ	13.46 kΩ	10 nF	68 nF	680 pF

The buffer has to be fast for this scheme to work correctly. When using a Bessel filter, transient response is obviously important. But if the buffer bandwidth is only 10× the corner frequency, 4% overshoot is produced. (25× the corner frequency gives 2% overshoot, and 100× gives 1%).

When using a Butterworth filter, monotonic frequency response is expected. But if the buffer bandwidth is only 10× the corner frequency, 6% peaking is produced. (25× the corner frequency gives 2% peaking, with 1% for a 40× ratio). For the Chebyshev, a 10× buffer bandwidth ratio gives 60% peaking. (25× gives 30% peaking; 100× gives 9%). This is a generic problem with the sharper roll-off filter characteristics. The components have to be more accurate and have better purities to maintain the performance. Low Q capacitors, for example, affect the Chebyshev more than the Butterworth or the Bessel types.

FIGURE 17.11H:



You make a high-pass filter by swapping over the capacitors and resistors. This means the capacitor values are no longer convenient. The practical solution is another table of values.

1 kHz (high-pass)	R1	R2	R3	C1	C2	C3
Bessel	8.356 kΩ	13.86 kΩ	44.02 kΩ	22 nF	10 nF	10 nF
Butterworth	6.054 kΩ	6.42 kΩ	48.61 kΩ	22 nF	10 nF	10 nF
0.5 dB Chebyshev	3.699 kΩ	3.844 kΩ	85.63 kΩ	22 nF	6.8 nF	10 nF

You will see application notes for standard Sallen-Key filters that show you how to wire the opamps up to give gain in the same stage as the filter. Realise that this would give even less buffer bandwidth and would be a very poor way of saving money since the filter response would be ruined.

Semi-conductor manufacturers now provide free software to design multi-pole active filters with various numbers of poles and filter characteristics. This gives a very quick and simple design path, including plots of frequency and time domain responses. Examples of useful free programs include:

FilterPro	**v2.00**	**Texas Instruments**	**www.ti.com**
FilterCAD	v3.0	Linear Technology	www.linear.com

Switched Capacitor Filters:

Setting up multi-pole filters to have the correct characteristic is not easy, as you have seen in the previous sections. The ideal would be to buy a monolithic part, which incorporated all the values. This would certainly reduce the amount of PCB space used. There are currently two ways that this can be achieved; one is a programmable analog filter chip, having resistor and capacitor values configured by programming the device. The other is a *switched capacitor* device, which gives very accurate ratios of the filter coefficients by adjusting the mark-space ratio of clocks inside the device.

Both of these technologies are relatively recent and are improving at a great rate. The programmable analog chip is of interest because not only can the corner frequency be programmable over at least a 10:1 range, but the characteristic can also be re-programmed during operation, according to the required function. This makes a versatile function, which would not otherwise be feasible in discrete components.

Switched capacitor filters are interesting for two particular reasons; one, an 8-pole filter can be implemented in a single chip and two, the corner frequency can be changed over a 100:1 range without any difficulty. The only thing you have to change is the clock frequency. This is a marvellous invention for those who need a variable filter function. This type of function would otherwise have to be done by sampling the data and using digital signal processing.

Thus, switched capacitor filters seem wonderful, but there are drawbacks which are not highlighted. *The switched capacitor filter has no attenuation at or near the clock frequency* or multiples of the clock frequency. In order to prevent this from causing an ***alias***, it is necessary to filter the signal before feeding it into the switched capacitor filter!

Currently available devices have a clock-to-corner-frequency ratio of between 140:1 and 10:1. If the corner frequency is 10 kHz then you might have a clock frequency of 1 MHz. The additional filter must be between 10 kHz and 1 MHz, must not attenuate the 10 kHz too badly, but must also attenuate the 1 MHz sufficiently. If the switched capacitor filter is an 8-pole filter then putting a single-pole filter at 100 kHz will not work. At 1 MHz the attenuation is only 20 dB, which is not sufficient, whilst at 10 kHz the additional attenuation is 0.5%.

It is therefore a balancing act as to the attenuation achieved at the clock frequency against the additional softness of the corner. Realistically a two or three pole anti-alias filter is needed. This is a real nuisance for clock tuneable filter applications, as discrete filter stages have to be switched in. The only good news is that a Butterworth anti-alias filter can be used in front of a Bessel switched capacitor filter, without adversely affecting the pulse response.

The other major problem with switched capacitor filter is harmonic distortion. Achieving better than –70 dBc distortion at tens of kilohertz is hard work. Such distortion levels are acceptable for 8-bit systems, but for 12-bit and higher the distortion is too high.

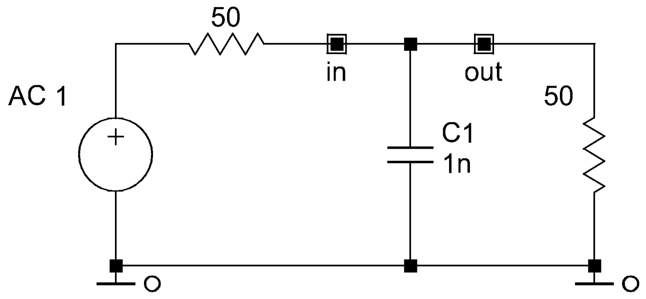
Matched Filters:

The previous filters have been for use in specific applications with low-impedance drives and high impedance loads. If the system is a matched transmission line then things change quite dramatically.

In a 50 Ω system, you ordinarily want components with 50 Ω input impedance and 50 Ω output impedance. This would ordinarily be a symmetrical component. Furthermore, it should ideally have a low reflection coefficient, usually expressed in terms of low VSWR.

Let's look at a simple low-pass filter in such an application.

FIGURE 17.11i:



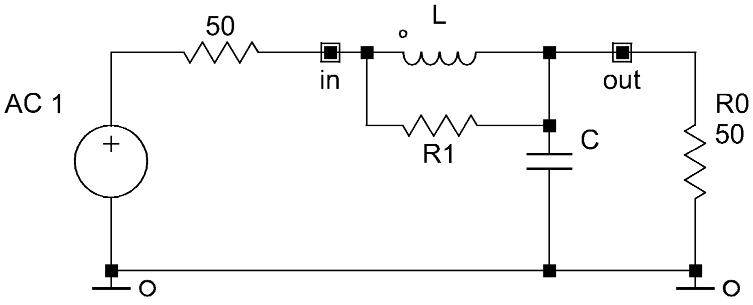
This low-pass filter consists simply of C1. It has several desirable features:

- ☺ Input/output symmetry.
- ☺ Low cost.
- ☺ Zero LF insertion loss.
- ☺ Low VSWR at LF.

However, as soon as the filter starts to filter, the VSWR becomes hopeless.

EX 17.11.4: What is the input VSWR of the above filter/load combination at the 3 dB bandwidth of the filter.

FIGURE 17.11J:



This filter is not symmetrical. However, it can be matched at all frequencies.

@EX 17.11.5: What is the relationship amongst C, L, R1 and R0 to give a VSWR of 1 at all frequencies.

You may not want to make matched filters, you may not have any interest in filters at all, but look at the circuit above more closely. If the signal source and the 50 Ω resistor represent a transmission line, and R0 is the receiving end with some stray capacitance C, then you can match the transmission line by adding an inductor in parallel with a resistor!

Digital Filters:

Books on digital filters [6] are widespread, large, and mathematically intense. It is also remarkably difficult to find the simple data that follows in these massive tomes. This is an introductory section for simple digital filter design, which can be supplemented by the specialist texts.[7] The first question to answer is: why would anyone want to use a digital filter? The answer is that once analog data is digitised, digital filters *are perceived to be* 'free', perfect, and infinitely variable.

Well nothing is free, but digital filters, when implemented in software for example, have no associated hardware cost. There is still a design cost to be considered, however, and there may also be a penalty in terms of the execution speed of the main system. Often, in order to get adequate performance, the digital filter has to be implemented as a separate floating-point digital signal processor (DSP), and this certainly will not be free.

Nothing is perfect and don't ever think something is! Digital filters do not suffer from harmonic distortion, change of characteristic with component values, or temperature related effects. However, digital filters come with their own new set of problems: round-off noise, aliasing, deep notches in the response characteristics, and lack of simple methods of obtaining filter coefficients.

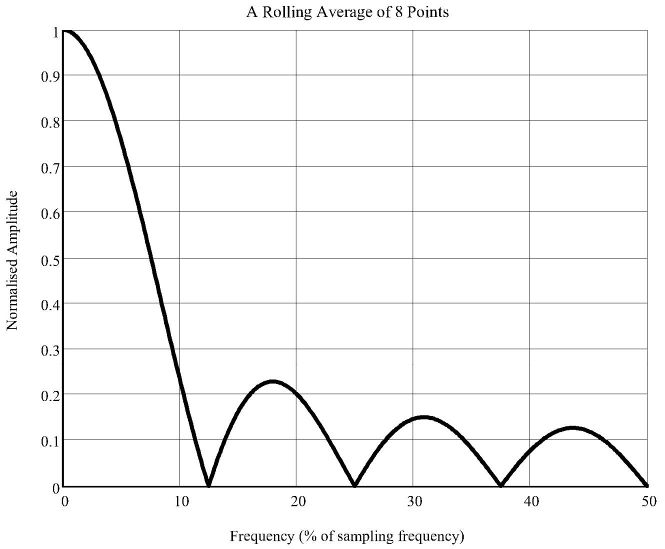
This section will only be concerned with low-pass filters. It is assumed throughout that the analog signal has been digitised using an ADC at a rate of F_S samples/second and the resulting data is processed in binary format. In this section a few simple filter designs will be introduced. If these are not suitable for your application then you will have to dive into a specialist book.

One simple digital filter consists of taking the mean of the last *n* acquired data points every time a new sample point arrives; this is referred to as a *rolling average*. As a hardware implementation you could think of the data being clocked from one data register to the next in a long chain, the output of each data register also being fed into an *n* word wide adder. If the current data word is W_1 and all previous data words up to W_n have been saved, the output word would be $W_{OUT} = \frac{1}{n}\sum_{p=1}^{n} W_p$. The 'divide by *n*' step can be eliminated if the words are stored in binary and *n* is an integer power of 2. In this case division is not necessary, the output register simply being 'tapped off' at the appropriate position. Suppose the data words are 8-bits wide. Summing 16 of these will produce a 12-bit result. You might decide to keep only the upper 10-bits of the result, increasing the resolution without using up too much memory. Nevertheless you have not actually had to do any division, it was merely a question of how the output register was connected to the adder output.

[6] R.W. Hamming, Digital Filters, 3rd edn (Prentice-Hall, 1989; repr., Dover, 1998).

[7] S.W. Smith, The Scientist and Engineer's Guide to Digital Signal Processing, 2nd edn (California Technical Publishing, 1999).

FIGURE 17.11K:



The plot of a rolling average of 8 sample points shows the key characteristics of this type of filter. For a block of n sample points, there are $n/2$ equally spaced nulls in the range up to 50% of the sampling frequency, F_S.

Operation beyond 50% of F_S causes aliasing and the frequency response characteristic gets mirrored about the 50% F_S line.

A simpler variant of this type of filter accumulates a block of n data points, outputs the result, then accumulates another set of n data points. The output data rate is reduced by a factor of n, but there is no need to have a large bank of n registers and a massive n-input parallel adder. The resulting filter function is identical to that shown above, except that the output data rate is n times slower.

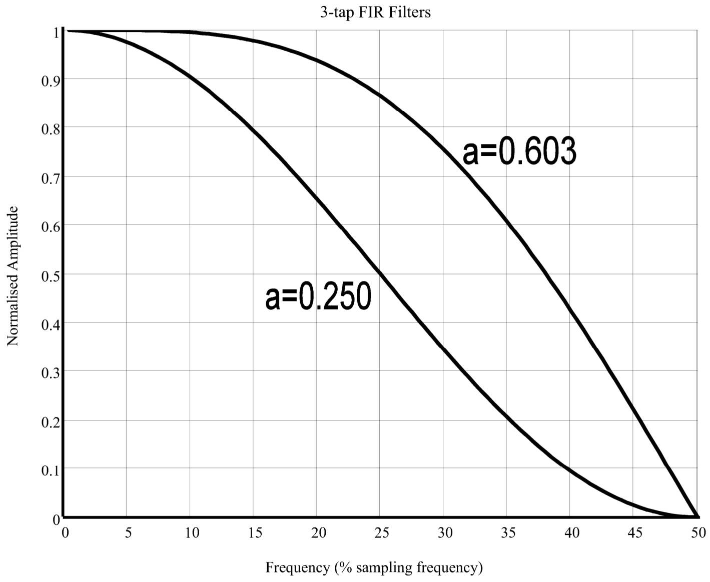
A more general type of filter function stores the n previous data points and sums them after applying a weighting factor to each data point. Suppose the last three data words are available, W_1, W_2, and W_3. The output data word is calculated from their weighted sum, $W_{OUT} = aW_1 + bW_2 + cW_3$. If it is decided that the low-pass filter should have a gain of unity at DC then $a + b + c = 1$. If the gain is also zero at 50% F_S ($b = a + c$), two of the three degrees of freedom have been used up. Then …

$$W_{OUT} = aW_1 + 0.50\,W_2 + (0.50 - a)W_3\,.$$

FIGURE 17.11L:

When a goes above 0.603 (or below –0.103) the response is no longer a monotonic decreasing function. $a = 0.25$ is particularly useful because it can be implemented without using a floating-point DSP.

A filter which only uses multiples of previous input data words is known as a *finite impulse response* (FIR) filter. It is also known as a *non-recursive filter*.



The general equation is $W_{OUT} = \sum_{p=1}^{n} a_p W_p$, where n is called the number of *taps*. Having more taps enables a faster roll-off filter to be created, but obviously the amount of hardware (or computational effort) is increased as a result. The array of weighting factors, [a_{p}], is also known as the *filter kernel*.

To implement high order multi-pole filters one could use a long series of taps, but the hardware (or computation) required for each input data clock would get excessive. In many applications it is therefore better to split the filter into smaller sections, making an 8-pole filter out of four un-equal 2-pole sections for example.

When the filter uses both multiples of the previous input data, and also multiples of previous output data, it is known as an *infinite impulse response* (IIR) filter, also known as a *recursive filter*. This type of filter can generate more useful filter functions with less hardware or computation.

The simplest and most widespread recursive filter is:

$$\boxed{W_{OUT} = (1-a)W_{IN} + a \cdot W_{OLD\_OUT}} \quad \text{where } 0 < a < 1\,, \quad a = \exp\left(-\frac{2\pi B}{F_S}\right)$$

In words: to get a 3 dB bandwidth of 10% F_{S} requires $a = 0.53$. This is achieved by taking 53% of the previous output and adding 47% of the new input. To see how the above equation for a is derived, consider an initial condition where both W_{IN} and W_{OUT_OLD} are 1. Then let W_{IN} drop instantaneously to 0 and consider the decay of W_{OUT} from 1 to 0. In each time-step, $\frac{1}{F_S}$, W_{OUT} drops by the factor a. Hence the output amplitude at time t after the input changed is $a^{t \cdot F_S}$.

Equate this decay to an equivalent exponential decay time-constant: $a^{t \cdot F_S} = \exp\left(-\frac{t}{\tau}\right)$

Since we know that $B = \frac{1}{2\pi \cdot \tau}$ for a single-pole low-pass filter: $a^{t \cdot F_S} = \exp(-2\pi B \cdot t)$

Take natural logs of both sides: $t \cdot F_S \cdot \ln(a) = -2\pi B \cdot t$

Divide by $t \cdot F_S$ and take the exp( ) of both sides to get $a = \exp\left(-\frac{2\pi B}{F_S}\right)$.

The subject of getting good coefficients for digital filters is so complex that a software package is typically required to calculate the coefficients for you. Fast roll-off filters can use hundreds or thousands of taps on the digital data path delay-line and hence hundreds or thousands of filter coefficients. High-order filters therefore need specialist DSP hardware to run at a reasonable speed, the design time being critically dependant on the available software tools.

17.12 Defensive Design

Defensive design has more to do with a philosophical point of view rather than any specific means of implementation. The key question you should be asking yourself is “What if…?”.

- ☹ What if the wrong value of resistor is fitted?
- ☹ What if there is a dry joint on this lead of this component?
- ☹ What if there is too much noise from the adjacent switched-mode supply?
- ☹ What if I need to probe this part of the circuit on the prototype?
- ☹ What if the manufacturer of this part decides they don’t want to make it any more?
- ☹ What if the test technician shorts out the power rail with his probe?
- ☹ What if somebody spills their coffee into the control panel?
- ☹ What if the mains lead gets pulled out accidentally when the equipment is being used?
- ☹ What if the user doesn’t read the user manual?

There is an infinite number of questions and you will not think of all of them. Just do the best you can. I obviously can’t answer every question that can be asked, but I can at least offer some solutions. Note that for any particular question there will in general be a multitude of possible answers. You have to think about the questions and the answers to decide if it is better to ignore that particular scenario, hoping that it will only happen rarely, or whether it would be wise to implement a “fall back position”. Some of the solutions are so easy that it is worth including them as a matter of course.

This list of solutions is entirely inadequate. Its primary purpose is to teach you *how* to think about this subject rather than *what* to think!

What if the wrong value of resistor is fitted?

If it is a component that only affects calibration then who cares? It will be picked up on test. If it is a power supply component, or an over-voltage trip component, then you have to say that the power supply should be tested before it is connected to anything important. Blowing up a $500 ADC because somebody fitted the wrong resistor value in a power supply is not very smart! The over-voltage trip would have to be checked by some sort of specific test.

What if there is a dry joint on this lead of this component?

Does the circuit destroy itself under these circumstances? If so, it is necessary to employ additional protection circuitry. It may be acceptable for a dry joint to cause mis-operation, but destroying other circuitry is unlikely to be acceptable. (This consideration is a typical requirement on the feedback circuitry around a switched-mode power supply.)

What if there is too much noise from the adjacent switched-mode supply?

Leave mounting holes for a screen over the switched-mode supply, or over the nearby circuitry, or both. Make provision for extra decoupling capacitors and or series resistors to kill off any conductively coupled noise. Split the power and ground planes, but allow them to be reconnected on the outside of the board to see which interconnection scheme is quieter.

What if I need to probe this part of the circuit on the prototype?

Add test points, including ground points for your scope probe. Remember that to probe a circuit without getting excessive noise you need a short ground lead to the probe. Thus it is useful to scatter lots of ground pickup points around a board so that the probe ground lead does not need to be 50 cm long.

What if the manufacturer of this part decides they don't want to make it any more?

Second source ordinary parts. On displays, power supplies, custom components &c you may not be able to second source the parts. Get assurances in writing of commitment to the part if possible. Otherwise check-out the track record of the supplier; do they make components obsolete without notice? Ensure an increased stock level of the part to allow for the re-design time.

What if the test technician shorts out the power rail with his probe?

Make the power supply short-circuit protected. Try to increase the spacing in areas where the circuit may be probed. Add deliberate test points so that the technician doesn't need to probe directly on the body of a surface mount package. Supply the technicians with fine tipped probes, rather than the robust long-life heavy-duty probes much favoured by managers and bean counters.

What if somebody spills their coffee into the control panel?

Make sure there are no high potential circuits under or near vent holes. Seal up entrances near where the user would be. Put louvers in the chassis which divert liquids into a defined safe area. Put a sign on the outside of the equipment "No food or drink in this area!" Use a conformal coat on the board so that it does not get affected by liquid, or dust accumulation.

What if the mains lead gets pulled out accidentally when the equipment is being used?

IEC connectors are designed to pull out easily to prevent accidents if somebody were to trip over the mains lead. Of course the mains lead should not be left where somebody could trip over it, but that is another matter. If the mains lead is accidentally pulled out then the ground {earth} for the equipment will also be lost. This could make the equipment case dangerous if the system is such that there could be ground current flowing. In this case an additional ground wire would need to be used to the equipment case.

What if the user doesn't read the user manual?

Expect the user not to read the manual and to do any required sequence of actions in some other arbitrary sequence. If this damages the equipment then expect lots of warranty returns, with the user insisting they did nothing wrong!

That was just my short list. I hope you got the idea.

CH18: temperature control

18.1 Temperature Scales

Temperature control is not a complicated subject, but simply because of lack of training on the subject, even apparently well educated engineers can have weird, and frankly stupid ideas about the subject.

It is of particular interest to be able to maintain an object at a constant temperature. The exact value of this temperature relative to national standards may not be critical, but constancy, despite the input power changing, internal power dissipation changing and the external ambient changing, is always important.

Weston Standard Cells are often temperature stabilised to better than 0.1°C, and then you have to apply a temperature correction to the voltage for the hundredths of a degree that the temperature is off from nominal.

Other examples of temperature stabilised items include metrology standard resistors, crystal oscillators, and zener voltage references. You always get the best possible device, in terms of TC, and then you make it a factor of 50× better (or more) by temperature stabilising the environment surrounding the device. It is not unheard of to do temperature stabilisation to 0.001°C and you have to appreciate that this will not happen by accident. You need to understand the subject very well.

There is a subject called *Control Engineering*. It applies to the stabilisation of mechanical servo loops for automatic controllers of various types. That it applies to the control loops of opamps and to the control loops of temperature controllers is not always appreciated {understood}. However, temperature control systems are quite different from opamp feedback schemes and radar tracking devices, so they are worth considering in their own right.

The quantitative definition of temperature is a huge subject, stretching back centuries. The first quantitative devices were gas thermometers, circa 1620, where the volume of gas enclosed at a constant pressure was used as the degree of hotness or coldness.[1] The type of thermometer using liquid sealed in a glass tube was developed around 1654, with mercury being used from around 1717. The Fahrenheit scale was created around this time, with the centigrade scale (developed by Celsius) following around 1744. The absolute temperature scale, based on the *second law of thermodynamics*, was proposed by Thompson (who later became Lord Kelvin) and quickly accepted by international agreement in 1887.

In the first instance the definition of temperature was somewhat arbitrary, the only important thing being the agreement between thermometers and the reproducibility of the scales. As science and engineering have progressed, it has become important to fix the scales relative to more fundamental phenomena. In other words, one finds an equation for the variation of something with temperature, then defines temperature by this equation. You can only do this trick once, so it is important to pick the right equation. All other equations and phenomena then have the possibility of non-linearities

[1] R. Weinstock, 'Temperature', *Encyclopaedia of Physics*, ed by Lerner & Trigg, 2nd edn (VCH publishers, 1991), pp 1246-1248.

with respect to this practical scale.

Certain easily reproducible physical phenomena are chosen as fixed points on the temperature scale; such phenomena naturally include melting points and boiling points. The remaining tasks are to define the position of these points on the temperature scales and also to define the method by which interpolation between the points is achieved. The result for the engineer is that the definition of temperature at resolutions of 0.01°C is not fixed for all time. The scale is periodically reviewed and adjusted to optimise agreement with fundamental metrology. The practical scale in use changed 8 times during the last century, and it is only to be hoped that such changes are less frequent in this century.

When you read in a book that the boiling point of oxygen is exactly −182.97°C, that was correct for the international temperature scale of 1927 (ITS-27) and 1948 (ITS-48), but was changed to −182.978°C for the international practical temperature scale of 1968 (IPTS-68) and does not appear as a reference point in ITS-90. The list of changes from IPTS-68 to ITS-90 temperatures [2] shows that the largest changes occur above 3500°C, the changes being around 2°C. Compare this with the change between the ITS-27 and ITS-48 values above 3500°C, which differ by more than 29°C.

Changing the temperature scales affects all physical constants, temperature coefficients and so on. Fortunately the scales have remained virtually unchanged in the 0°C to 100°C range for the last hundred years. The 1990 changes coincided with the changes to the fundamental electrical units, requiring only one new set of values to use. The fundamental metrology has now reached such a high level of precision that future changes are unlikely to have much impact outside of national metrology institutes.

18.2 Precise Temperature Measurements

It is instructive to look at the errors that occur with precision mercury-in-glass thermometers. The errors that occur are easy to visualise and the same type of errors should be expected for all temperature measurement systems, although the physical mechanism of the errors will clearly be different.

Mercury-in-glass thermometers are available for specific temperature ranges with excellent accuracy and resolution. For example, the thermometers used on a particular brand of temperature stabilised standard cells had scale lengths of about 4 cm, calibrated from 37.00°C to 38.00°C. You may have supposed that having such a thermometer, complete with a calibration certificate, would be all you needed to make an accurate measurement of temperature. For an accuracy of ±0.1°C you would probably be correct, but for an accuracy of better than ±0.01°C you could certainly be wrong.

A mercury-in-glass thermometer measures the differential expansion between the glass and the mercury. Therefore the temperature of the stem of the thermometer affects the reading. A first order correction for the stem being at a different temperature to the bulb is: $0.00016 \times \Delta T_{SCALE} \times (T - T_{STEM})$ °C, where: ΔT_{SCALE} is the exposed length of the mercury column in °C, T is the measured temperature, and T_{STEM} is the mean stem temperature.[3] The generic name *immersion depth* is used for any type of sensor probe

[2] B.W. Mangum, and G.T. Furukawa, 'NIST Technical Note 1265', *Guidelines for Realizing the International Temperature Scale of 1990 (ITS-90)* (National Institute for Standards and Technology, August 1990).

[3] National Physical Laboratory, *Calibration of Temperature Measuring Instruments*, 3rd edn (HMSO, 1964), Notes on Applied Science No. 12.

that can be either fully or partially inserted into the sensed region.

The next error to watch out for is offset drift, sometime referred to as ‘zero drift’. The bulb containing the mercury is made of glass. Glass is not a solid, but a super-cooled liquid, and it creeps with time. The bulb will gradually contract over a period of years causing an effective offset on the scale, a ‘zero error’. Thus routine annual calibration is important, certainly for the first few years of a thermometer’s life.

Another interesting error is due to the rate of change of temperature. If the thermometer was at say 100°C, then was cooled to say 0°C, the initial readings could be lower than 0°C by ≈0.05°C. This error settles out over a period of *hours*, depending on the exact type of glass used. Replacing the glass by fused quartz can reduce this *zero depression* effect by more than an order of magnitude.

Atmospheric pressure affects the reading on a mercury-in-glass thermometer, but perhaps less obviously, the immersion of the thermometer into a liquid also affects the reading. Suppose you insert the thermometer into the bottom of a tank of water, or worse still, some more dense liquid; you now have a significant pressure change to take into account. Oh, and did you allow a correction for the angle of the thermometer relative to the vertical?

The key point of this section was to force you to think of sources of error. If you now measure temperature, or any other quantity for that matter, hopefully you will consider the environmental factors that could increase your measurement uncertainty from the values given on the calibration certificate.

In your career as a designer you should be creating new methods, techniques and equipment. These new things may suffer from a whole battery of new effects which nobody has yet thought of. Being in the lead, you will need to remember past problems to see where new problems may be.

18.3 Temperature Sensors

The first thing you should be aware of is that thermocouples are very good for measuring large ranges of temperatures. Secondly, they are physically very small (≈1 mm diameter) so that they can measure small objects without introducing large time-constants or changing the temperature significantly. A slight problem with thermocouples is that they are *differential* temperature devices; they only measure the difference in temperature between one junction and the other. There are *always* at least two junctions, although as a user of hand-held thermocouple meters, you might be forgiven for not realising that the meter itself contains one of the junctions.

In common use you plug a two-terminal thermocouple into a hand-held electronic thermometer and you ‘know’ there is only one thermocouple junction. This is a convenience for the user and is not part of the engineering phenomenon of the ***Seebeck Effect***. These hand-held digital meters are so inexpensive, and so widespread, that a whole generation of engineers may forget their basic physics and not realise that the second junction [the *reference junction*, often called the *cold-junction*] is formed at the connector, or just inside the meter body, depending on the exact internal construction.

If there is no temperature difference between the ‘real’ thermocouple {the part you use as a probe} and the body of the meter, then there will be no voltage generated by the thermocouple. Now I have had “engineers” argue with me on this point, since the meter clearly reads the current ambient temperature. Ah! That is because the meter has a built in ambient temperature sensor known as the *cold-junction compensation*. If the thermocouple probe is at the same temperature as the thermometer body, all you are reading is the accuracy of the cold-junction compensation circuitry.

Why is this important? If you decide to measure an ambient temperature with your hand-held meter by carrying it into an area with a different ambient temperature, you will initially get an incorrect reading! The bare thermocouple will respond to the new temperature almost immediately (seconds) whereas the cold-junction compensation in the meter may well take up to several minutes to settle down at the new ambient temperature.

If you wanted to do an accurate job and not use a cold-junction compensator you would need a different sort of digital thermometer; one that had a deliberate external reference junction. This reference junction could then be held at some convenient temperature, by immersion in an ice-bath for example. You would then get a very accurate measuring device which was relatively unaffected by ambient temperature changes.

It is possible to get an accurate and stable cold-junction compensator, but it is evident that an additional uncertainty is being entered into the measurement. If you can measure the cold-junction accurately by some other means, then you can use the same method to measure the ambient temperature in the first place; clearly the direct measurement will give less measurement uncertainty.

Thermocouples are often far away from the measuring device. Rather than use expensive thermocouple wire to cover the distance, it is usual to extend the thermocouple leads using special wire. *It is absolutely vital that copper wire is not used*. If you use copper wire to extend a thermocouple, you get an extra pair of thermocouple junctions formed out in the open in some uncontrolled environment. Extension cable is made from the same alloys as the thermocouple, so no additional thermocouple junctions are formed, although it can be expensive. The joints also have to be made between identical alloys, another way to inadvertently ruin the measurement accuracy! Connect the extension cable backwards and you create more junctions and get incorrect readings.

A cheaper solution is to use *compensating cable*. This needs to be the correct type of compensating cable for the thermocouple you are using. The alloys chosen give low Seebeck coefficients for the joints and thus minimise the ambient variation errors. Obviously compensating cable also needs to be correctly polarised relative to the thermocouple.

For industrial use, thermocouples are usually considered as completely *interchangeable*. What this means is that when you replace a thermocouple by another of the same type, the calibration is unchanged. This is obviously not a totally true statement; it is an approximation that is workable to better than a few degrees Celsius. Although the interchangeability of the thermocouple may leave the particular thermometer still within its stated uncertainty, that doesn't mean that no interchangeability error will occur. Since thermocouples are made of alloys, the exact composition will affect the calibration. If you dig a bit deeper you will find international standards that give tables of thermocouple voltages and error limits.

IEC60584:1995 gives tables of output voltage for various thermocouple types and also gives various accuracy classes. To be specific, take the example of the popular type-K, Nickel-Chrome / Nickel Aluminium thermocouple. The thermocouple tables are generated from 10th order polynomials, with the coefficients given to 11 decimal places. In the temperature range from −40°C to +333°C a tolerance class 1 type-K thermocouple is accurate to ±1.5°C. A tolerance class 2 type-K thermocouple is only accurate to ±2.5°C.

If you design your own thermocouple signal conditioner then here is a tip. For safety reasons it is important that if the thermocouple breaks the controlled system does not melt as a result. Protection is easily accomplished by using a single extra resistor. An ordinary thermocouple junction, even at the ends of long leads should not exhibit more than say 1 Ω of resistance. Suppose it is driving an amplifier with an input impedance of 1 MΩ. If the thermocouple breaks the temperature indicated would be ambient since there would be no voltage developed. The temperature control system would then apply full power indefinitely.

A well designed rapid-response temperature control system would have lots of spare power in hand so that the initial rate of rise of temperature at power-on could be minimised. The result is that full power applied indefinitely would probably melt something. The simple answer is to connect a high value resistor from a power rail to the thermocouple. Suppose you connect a 10 MΩ resistor between the +5 V power rail and the thermocouple input. In normal use you would get an error voltage of much less than 0.5 µV. This is not much of an error, and could in any case be calibrated out.

When the thermocouple breaks, or becomes disconnected, the 10 MΩ resistor pulls the input up to 455 mV. This high a voltage will undoubtedly trip the internal over-temperature cut-outs within the controller and safely shut the system down. As a designer you should regard this as a minimum design spec, but as a user you should check to make sure that the controller you are using actually does this. Don't be an ***ought-to*** engineer! Actually, any well designed system would also have an over-temperature cut-out in case the primary control system failed. This should be as independent of the main system as possible, preferably having its own temperature sensor.

For general temperature sensing applications there are several inexpensive resistive devices available. Least expensive is the NTC thermistor (**N**egative **T**emperature **C**oefficient). The value of resistance *decreases* with temperature according to a logarithmic curve. The slope of this curve and its initial value both have wide tolerances (between ±2% and ±20%). They are good for general purpose temperature control or measurement applications over the range of –10°C to +100°C (different thermistors are optimised for different bands of temperature). What is not specified is the stability of the measurement with time. Being resistive devices one should expect that the device would drift with time to some extent. I would not want to rely on this stability much below the level of ±3°C. Interchangeability of thermistors is generally very poor and recalibration is usually necessary.

Another type is the PTC thermistor [**P**ositive **T**emperature **C**oefficient]. This type is interesting because it can be used for temperature control applications on its own. The PTC curve has a sharp 'knee' in its curve at a particular temperature. This can be used as a cut-out or control signal. PTC thermistors are often specified at this knee temperature, which would probably be within a 5°C band. Again I wouldn't want to rely on the stability of this type of device to better than ±3°C without manufacturer's data to support the stability claim.

Interpolation between the fixed points of the international temperature scale (ITS-90) is done using a standard platinum resistance thermometer between 13.8033°K and 1234.93°K. The coefficients used in the interpolation formula change according to the sub-range being measured, but the point is that the corrected platinum TC curve is being used to *define* temperature. It should therefore be evident that a platinum resistance thermometer can give the most accurate reproduction of the temperature scale.

The Platinum Resistance Thermometer (PRT; or Platinum Resistance Device) is a long thin winding pattern of platinum film, usually on a ceramic substrate. A common type, PT100, has a 100 Ω resistance at 0°C. Unfortunately you often see them sold as part of a two-wire probe assembly. This just ruins the performance of the device in the terms of accuracies to small parts of a degree because of the resistance of the copper leads and the junctions formed with the platinum element. Obviously a proper 4-wire termination to the platinum element is highly desirable.

Typical error sources when using a PRT:

a) Using the wrong corrections factors for the type of PRT you are using.
b) Too much current in the PRT giving a self-heating error (use 1 mA or less).
c) Thermal EMFs in the lead wires.
d) Lead resistance due to a 2-wire connection to the PRT.
e) Heat flow through the lead wires making the PRT cooler than the sensed region.
f) Inadequate immersion depth.

LAW: A thermometer only senses *its own* temperature.

It is possible to use base-emitter junctions as temperature sensors, using the simple rule that the voltage across a forward biassed silicon base-emitter junction changes by approximately –2 mV/°C. This can give an inexpensive temperature sensor, but the accuracy is far from certain. Base-emitter TCs can range from –1 mV/°C to –4 mV/°C, so some slope and offset calibration would certainly be necessary, and this calibration cost needs to be considered when comparing the overall sub-system cost. Nowadays it is convenient to buy a complete ‘digital’ solution to the problem, where the temperature sensor, linearisation and interface are all built into one inexpensive chip. This integration saves a huge amount of design time and PCB area, so it is well worth considering for routine monitoring of internal instrument temperatures, for example.

18.4 Heat Transportation

The mechanisms of heat transport are well known and should be familiar to you.

@EX 18.4.1:

a) What are the *four* modes of heat transportation?
b) What is the dominant mode in the region of 20°C to 120°C?

Heat transportation is of great interest to electrical engineers because excess temperature causes premature failure of electrical and electronic components. The number of learned texts on the subject of heat transport is large and the mathematical complexity is great. Thus, whilst you are acutely aware that there is a complex problem to solve, you only have a few minutes in a day allocated to handling these 'trivial' problems.

Obviously the first idea is not to generate excess heat. That means higher efficiency. However, higher efficiency costs more by increased design time and possibly a higher unit cost. In terms of tangible costs, it is often necessary to dissipate a bit more heat than is absolutely necessary. This might mean settling for 85% efficiency, when a circuit with 90% efficiency would take twice as long to design.

Dissipating heat is seen as a "cheap" option, as the depletion of natural resources is not visible to accountants. Also, if a device gets too hot then you just put on a bigger heatsink or a bigger fan. This action tends to be very empirical {experimental} and is therefore neglected by 'academic' textbooks.

18.5 Thermal Models

In the real world, at temperatures in the region of 0°C to 120°C, heat transport mechanisms can be considered to be *linear*, whereas physics textbooks give radiated heat transfer as a forth power effect. For now, just ignore the fourth order powers; the result is a safer analysis in that the device will run cooler than expected. Just lump all the cooling mechanisms into a single linear factor. If the device is running at 60°C in an ambient of 20°C then it is predicted that at an ambient of 50°C the device will be running at a temperature of 90°C. That's it; that's the majority of what you need to know!

Now if the thing you are going to build is expensive to physically model, you will want to do some maths first, saving the cost of making several models. Otherwise, just 'guess' what cooling you need and as you get better at guessing you will be a more experienced (and valuable) engineer. One way to improve your guessing is to look in manufacturer's catalogues of heatsinks. They give thermal resistance (°C/W) figures for their heatsinks and from these you will be able to see how much cooling a certain amount of fin area gives.

Thermal Resistance is a very simple idea, defined as follows:

$$\boxed{\text{heat flow} = \frac{\text{temperature difference}}{\text{thermal resistance}}} \qquad P = \frac{\Delta T}{R_{TH}}$$

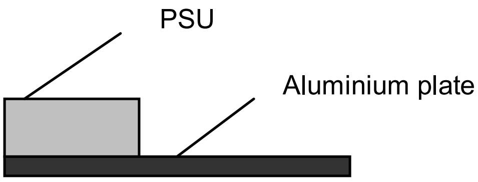
The temperature difference is in degrees Celsius; the heat flow is in Watts. Hence the thermal resistance has units of °C/W. The equation has the same form as Ohm's law, with thermal resistance equivalent to electrical resistance, temperature difference equivalent to potential difference, and heat flow equivalent to electric current. Paths in series therefore add their thermal resistances, whilst paths in parallel combine in the same way as paralleled resistors (reciprocal of sum of reciprocals).

***EX 18.5.1:** A power MOSFET has a thermal resistance junction-to-case ($\theta_{J\text{-}C}$) of 2°C/W. It is mounted onto a heatsink with a thermal pad specified at 1°C/W. The heatsink is rated at 10°C/W under the air flow conditions in which it is being used.

a) If the MOSFET is dissipating 8 W and the ambient is 50°C what is the junction temperature?
b) In another mode of operation, the heatsink is at 71°C in a 20°C ambient. How hot is the MOSFET junction {die temperature}?

Heat conduction is a very easy mechanism to both understand and to do calculations with. You have to be able to do "back of an envelope" calculations of simple problems, rather than resort to full blown analyses; there is usually not the time to do anything else. If you are working on a prototype that is going to cost $900,000 or more, it may be worth investing the time in doing more rigorous calculations. However, for ordinary applications where you can build a model relatively cheaply, excessive calculation is inadvisable.

EX 18.5.2: A power supply is bolted to an Aluminium chassis plate to cool it. The power supply is delivering 100 W and the manufacturer states the efficiency as ≈70%. There is no fan in the system and the power supply is in good thermal contact with the chassis plate.



The Aluminium chassis plate is 2 mm× 140 mm× 210 mm. Is the plate thick enough to spread the heat effectively? Take the thermal conductivity of aluminium as 230 W/(m·°C). *Hint: Don't try to calculate this accurately.*

Convective heat transfer is the usual way of finally getting rid of heat to the ambient. Natural convection has heat rising away from the hot body and cooler ambient air moving in from underneath to fill the gap. The standard idea when using pumps or fans is to drive the air in the direction it is naturally going in; working *with* nature rather than against it.

The formula for convective heat transfer [given by Newton in 1701] is:

$$\boxed{\textit{heat transfer rate},\quad P = h \times A \times \Delta T}$$

This formula relates the heat transfer rate (watts) to the surface area (m^2) exposed to the 'fluid', the temperature difference between the heated object and the fluid, and a 'constant' *h*. The problem with this neat little formula is the value of *h*.

h is actually a function of the characteristic length of the surface, the type of fluid, the flow rate of the fluid and the temperature! Thus *h* is a parameter, rather than a constant, and it ranges from below 0.6 W/(m²·°C) to above 7 W/(m²·°C) just for natural convection in air.[4]

[4] J.R. Simonson, 'Approximate Formulae for Use with Air' in *Engineering Heat Transfer* (Macmillan Press, 1975), pp. 120-122.

H and *L* are height and length respectively, measured in metres.

h in W/(m^{2}·°C)	**Laminar flow**	Turbulent flow
One side of a vertical surface	$h = 1.41 \times \left[\frac{\Delta T}{H}\right]^{0.25}$	$h = 1.31 \times [\Delta T]^{0.33}$
Horizontal surface facing up	$h = 1.31 \times \left[\frac{\Delta T}{L}\right]^{0.25}$	$h = 1.52 \times [\Delta T]^{0.33}$
‘Horizontal surface facing down	$h = 0.58 \times \left[\frac{\Delta T}{L}\right]^{0.25}$	

- ✓ Turbulent flow is predicted when $L^3 \times \Delta T > 1.56$, where L is the characteristic linear dimension of the system in metres. [For ΔT > 15°C, L has to be greater than 0.47 m for turbulent flow.]
- ✓ A vertical surface may get converted to a horizontal surface if a piece of equipment is standing on its back feet.
- ✓ For a given surface area, it is better to have a small feature size ie lots of little fins. However, the improvement is very ‘slow’, so that to double the cooling, the features have to be reduced by sixteen times the linear dimension.

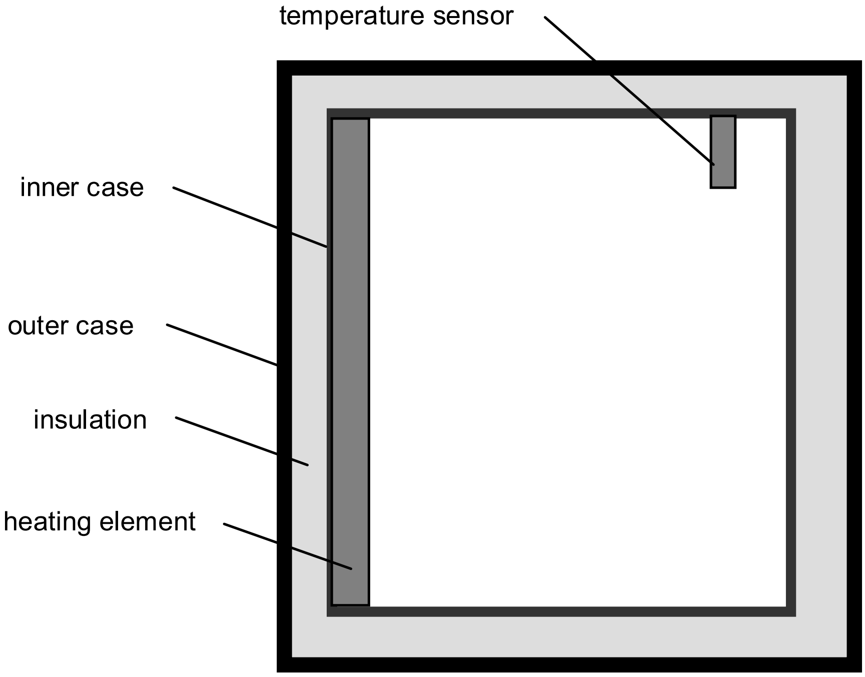
The thermal resistance is given by

$$\boxed{R_{TH} = \frac{1}{h \times A}}$$

Now that was a steady state problem, but dynamic situations also occur. For this you need to look at heat capacity as it relates to heat transfer. From your physics courses you should remember *specific heat capacity* as the heat energy (joules) needed to raise the temperature of a 1 kg block of the material by 1°C. Thus for a particular piece of material there will be a heat capacity in J/°C. This means that when heat flows through a material there is a steady state solution which involves thermal resistances, and there is a transient solution which involves heat capacity. Electrical capacitance is the equivalent of heat capacity in the model being used.

Consider a simple temperature controlled laboratory oven.

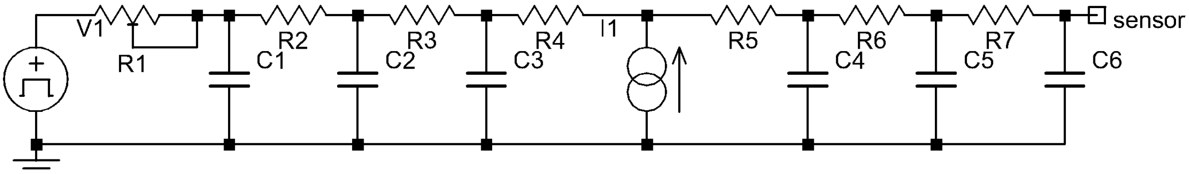
FIGURE 18.5B:



Notice that the heating element goes closer to the outside ambient than the thing being controlled. This is a fundamental of temperature control. You are trying to make a temperature controlled space {area; region}. What you do not want is a temperature gradient in this space. *Steady state heat only flows to the outside ambient.* Heat does not flow through the controlled space, thus there is no temperature drop across the space.

It is important to understand this principle when designing temperature controlled rooms. Heat is lost by the walls (in the case of a room this includes the floor and ceiling). Heat is also lost via the air replenishment system. Heat is gained by electrical equipment and generated by human bodies (≈100 W resting, and 300 W to 500 W for light work.)

FIGURE 18.5C:



Working left to right: V1 is the ambient temperature; R1 is the (variable) thermal resistance from the case to ambient; C1, C2, C3 and R2, R3, R4 are a simplified equivalent of the insulation; I1 is the power in the heater element; R5, R6, R7 and C4, C5, C6 are a simplified equivalent of the heater-to-sensor thermal path.

You should appreciate that the environment has two variables to adjust; V1 and R1. In other words the ambient temperature can change and the thermal resistance from the ambient can change. The thermal resistance to ambient will almost certainly be dominated by convection cooling. If there is a draft, or a howling gale, then the heat exchange rate will be increased over that found in still air.

The normal operating principle of such an oven is that the ambient changes are "filtered out" by the insulation time-constants; the heater and sensor do not get exposed to these excursions to any great degree. This idea is acceptable if you are only looking for a few °C control accuracy, but is not acceptable when you need <0.01°C stability.

18.6 Open-Loop control systems

You might not think there was such a thing as an *open-loop controller*. After all, if there is no feedback, how can control be established? Well, people have been successfully boiling eggs in saucepans for a great many years without the cooker sensing the temperature of the saucepan or its contents. The control is a power control and any feedback is only by human intervention.

Actually boiling water is a poor example; any school kid knows that water boils at a constant 100°C (unless you change the atmospheric pressure or add impurities). Nevertheless, any adult should be able to correctly set the power level to gently heat baked beans. This is possible because during the cooking period, the thermal resistance from saucepan to ambient is reasonably constant, and the ambient is not changing dramatically either.

Another example of open-loop control is a simple shower unit. The water temperature is adjusted by changing the flow rate. Often this is supplemented by a coarse power control in perhaps two or three steps. Strictly speaking these are in fact closed-loop control systems, with the operator being part of the feedback loop.

18.7 ON/OFF temperature control systems

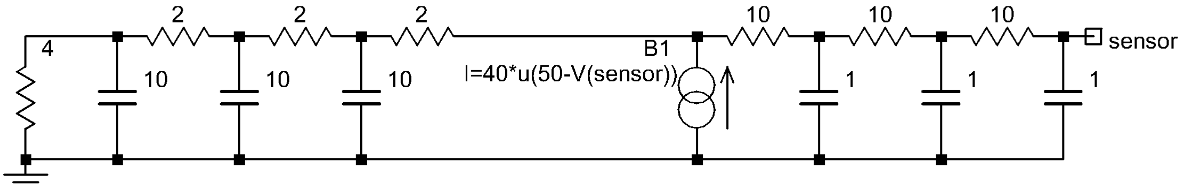
As far as a closed-loop control system is concerned, the simplest type is the on/off controller. You can look at this as a *thermostat*. If the temperature is too high the thermostat switches the heater power off. When the temperature drops, the thermostat turns the heater power back on. The thermostat switching points are always separated by at least a few degrees. This inherent hysteresis on the switching point is fundamental to the design of a mechanical thermostatic switch. Some therefore mistakenly believe that replacing the mechanical thermostat with an electronic equivalent having no hysteresis at all will make a good control system; they need to study the subject more fully!

The erroneous idea that zero hysteresis means no temperature excursions probably stems from simplifying the system down to a single RC time constant. I have deliberately shown the system as having three poles to avoid this trap. A single RC system with no hysteresis would indeed exhibit a minimal temperature excursion. However, in the real system the temperature must always overshoot because of the inherent delay in the sense path.

If the power in the heating element is only slightly higher than the amount needed to sustain the inside at the required temperature, the oven will take far too long to reach the required temperature. Hence the heater needs to be at least twice as powerful as that, if not five times as powerful. This “overcapacity” allows the oven to heat up faster. The switch-on temperature response of the heater may then appear as a simple ramp, rather than a complex exponential curve. It *is* exponential, but the first part of an exponential waveform looks linear, and that is the only part which will be seen. Because the sensor is remote from the heater, this temperature ramp is only detected some time later. There is therefore excess heat (and temperature) which needs to dissipate before the power can be turned back on.

The behaviour is very easy to simulate.

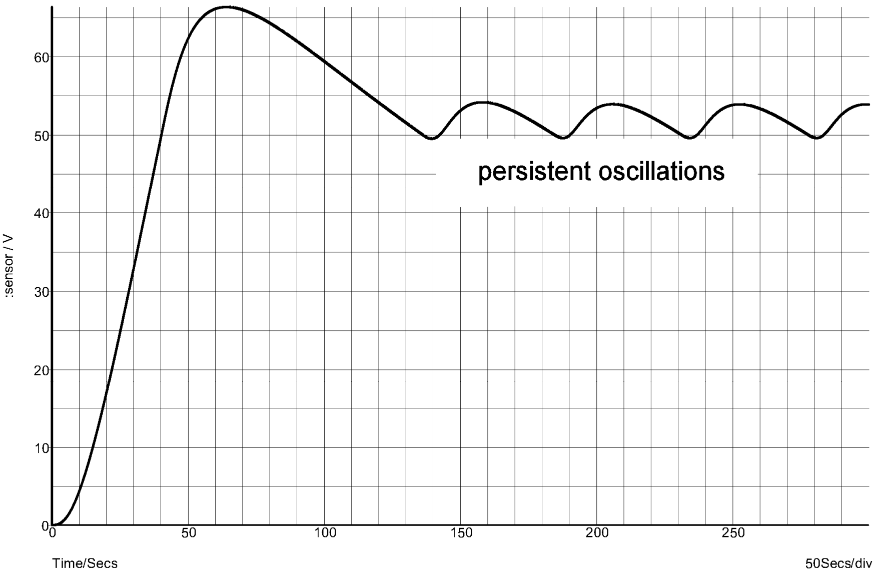
FIGURE 18.7A:



The arbitrary current source B1 is the key to the simulation. The function u(x) returns 1 for $x>0$ and 0 for $x<0$. This model accurately simulates a zero-hysteresis on/off control system.

FIGURE 18.7B:

The soft initial corner is found on all real systems due to the distributed time constant. The first overshoot is due to the high *aiming value*; the high aiming value being produced by the overcapacity in the heating system required for a fast warm-up.



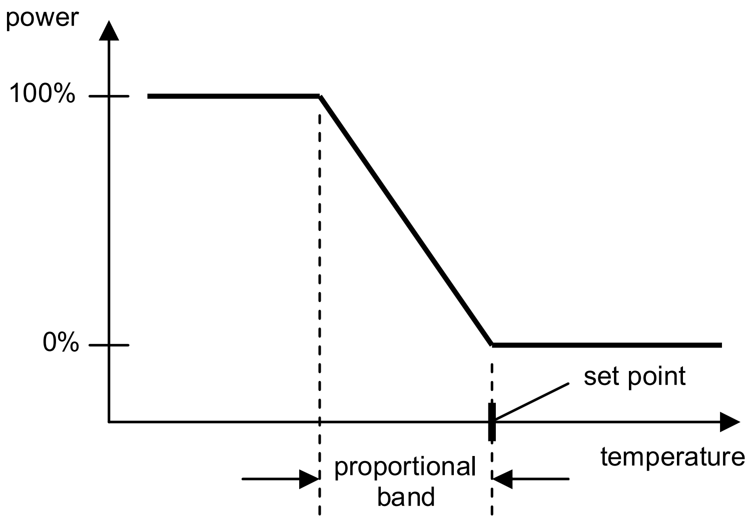
The warm-up time of a given physical system is initially dependant only on the power in the heater and has nothing to do with the control system. The fact that the ripples are all above the set temperature is just a question of calibration. It is easy enough to adjust the calibration so that the set temperature is more evenly distributed about the ripple band. Nevertheless it should be clear that even a zero hysteresis on/off controller gives temperature fluctuations.

Having seen what causes the temperature fluctuations, you may be tempted to put the sensor right next to the heater. This would be a poor choice because the temperature of the heater varies dramatically according to the applied power. Another "idea" would be to use more insulation between the controlled item and the sensed point. It is true that this would tend to filter out the temperature excursions, but what then happens is that the sensor cannot respond to power changes in the unit being temperature stabilised. This gives another a poor temperature control system.

18.8 Proportional control

Notice that those output excursions have nothing to do with ambient changes. In the simulation the ambient temperature is constant. In order to avoid the ripples you could use a *proportional* control system. This is a very simple concept; the power applied is proportional to the difference between the set point and the actual temperature over a limited band of temperatures. This is best illustrated with a diagram.

FIGURE 18.8A:

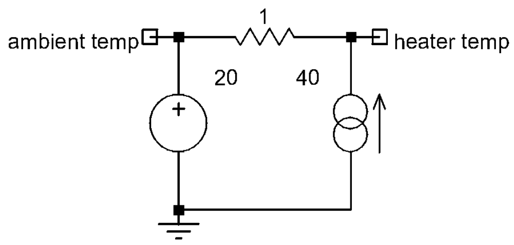


It is initially convenient to define the set point as the 0% end of the proportional band. This means the steady state temperature will always be lower than the set point, but it should be well within the proportional band.

Full heater power must take the temperature well beyond the set point in order for the control system to work correctly.

The proportional gain factor is set by adjusting the width of the proportional band. This is perhaps an unfortunate term; you end up adjusting the proportional-band width. In spoken usage this can easily be mis-heard as the 'proportional bandwidth' and then confusion can occur.

FIGURE 18.8B:



I have (arbitrarily) to referenced temperatures to 0°C. 1 V represents 1°C, 1 A represents 1 W and 1 Ω represents 1°C/W: this is a simple steady-state model. In the steady state, the heater temperature and the sensor temperature are equal (according to the previous model, where there is no thermal path from the sensor to ambient).

Using symbols:

T_{SET} = set temperature
T = actual temperature
T_{AMB} = ambient temperature
T_{BAND} = proportional-band width
P_{MAX} = maximum possible heater power
$\theta_{H\text{-}A}$ = thermal resistance from heater to ambient

$$heater\ power = P_{MAX} \times \frac{T_{SET} - T}{T_{BAND}}$$

Using the steady state model:

$$T = T_{AMB} + (heater\ power) \cdot \theta_{H-A}$$

$$T = T_{AMB} + P_{MAX} \times \frac{T_{SET} - T}{T_{BAND}} \times \theta_{H-A}$$

Simplify this by writing $K = \dfrac{P_{MAX} \times \theta_{H-A}}{T_{BAND}}$ giving $T = T_{AMB} + K \times (T_{SET} - T)$

Then $$T = \frac{T_{AMB} + K \cdot T_{SET}}{1+K} = \frac{T_{AMB}}{1+K} + \frac{T_{SET}}{1+\frac{1}{K}} = \frac{T_{AMB}}{1+K} + T'_{SET}$$

Where T'_{SET} is the re-calibrated set temperature, taking into account the $\frac{1}{1+\frac{1}{K}}$ factor.

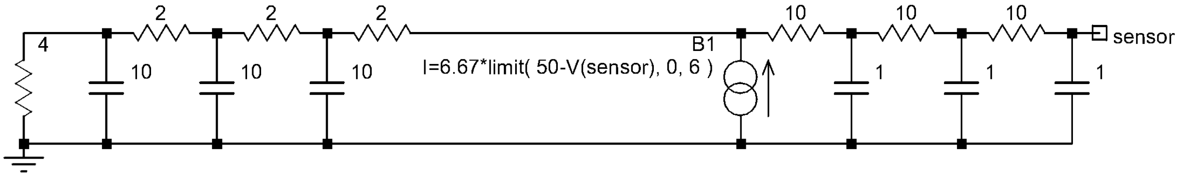
This equation is valid only if $\theta_{H\text{-}A}$ is relatively constant.

There is an ambient temperature desensitisation factor of $\dfrac{1}{1+\dfrac{P_{MAX}\theta_{H-A}}{T_{BAND}}}$.

Since you don't want the controlled temperature to change with changes in the ambient you would think that you should just make this term tend to zero. If this is done then you end up with an on/off control with zero hysteresis, which you know produces output temperature ripples. In other words if the proportional gain is made too high, by making the proportional-band width very small, then oscillation results.

The instability is caused by the multi-pole response of the thermal insulation between the sensor and the heater.

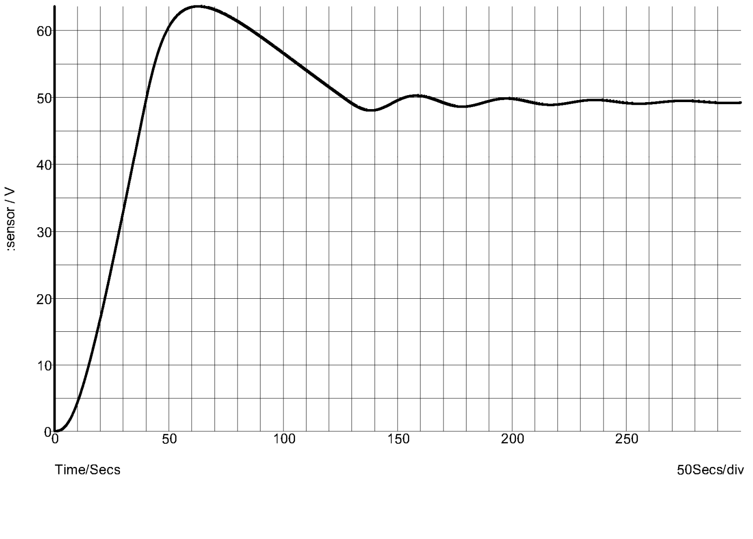
FIGURE 18.8C:



The function **limit(X, L, H)** returns L for X<L; H for X>H; and X otherwise.

This system still shows overshoot and ringing because the proportional band is too narrow. Reducing the time constants from the sensor to the heater seems like the correct solution. However, in this simple model there is no 'load'. Imagine that the oven is trying to heat up something like a piece of equipment. This has its own heat capacity and you want to control its temperature not just the temperature of the heater. The correct solution is to use a fan to blow the air around in the oven. This shortens the time constants and therefore speeds up the response of the system.

FIGURE 18.8D:



All the usual control system theory can be applied to this system to make it stable. One can draw out a Bode plot of the response and use dominant pole compensation to stop the multi-pole sensor response causing oscillation. The next steps would include adding an integral term to remove the long term error and adding a differential term to remove the overshoot.

This gives what is known as a *three-term control system*. This is the 'classic' way of getting a control system to work. The coefficients of the three terms are adjusted and then fixed for an optimum response. Modern controllers can come with software which automatically calculates the required coefficients, greatly minimising the time required to tune the system response.

One of the fundamentals of analog engineering is to do as little as possible of it. Long time constants (>1 s) in analog systems are very prone to noise, drift and difficulty of adjustment. By contrast even the most inexpensive of low-speed digital circuitry can handle this speed of response with ease.

Now I am not proposing that all temperature control systems should be replaced by digitally controlled schemes. Use common sense and use simple analog schemes for simple problems. When the thing you are controlling is expensive and you want to optimise the response then put in a digital scheme.

The digital scheme could just emulate the analog performance. Alternatively the system could dynamically adapt to changing conditions, a much more powerful control strategy. 8-bit micro controllers with built in ADCs are cheaply available (≈$3.00) and these can implement such control systems with ease. Expect there to be significant cost in developing the software however.

Proportional power is not a difficult thing to deliver. The traditional way of doing it is to use an on/off switch cycling at a rate of say 30 seconds. The ***duty cycle*** is then a proportional control of the power. This slow cycle rate is why relays can be used to provide proportional power without burning out due to excessive switching. In fact this is also how conventional heated ring electric cookers still work today. In the trade the (open-loop) control is called an *energy regulator* and consists of a heated bimetallic mechanism, biassed by a cam to change the duty cycle. It is cheap, reliable and proven technology.

This method of providing proportional power is efficient because there is very little power loss in the switching device. It is an interesting point of scaling that bigger systems, which necessarily require more power, have longer time-constants. It is therefore often not a problem to have the switching device cycling with a 30 second period. Obviously if the switching device has to cycle at <1 s, an electronic switching device will be necessary. Notice also that these types of control systems are now potentially problematic due to EMC flicker requirements. If you are taking 4 MW bursts of power every 10 seconds, that is likely to make the lights pulse!

Another method of delivering proportional power in the past has been a *phase controller*. On an AC supply, a thyristor would be fired at some point after the zero crossing point. Adjustment of the phase delay gave an inexpensive and efficient power control scheme. This scheme has been limited in application by EMC regulations since it generates so much RF interference and harmonic currents. If the full load is ≤ 100 W then it is acceptable; otherwise there is a limited number of cases where phase control is acceptable. See EN61000-3-2:2000.

The 'burst firing' scheme overcomes the emission problems of the phase controller by always switching the power on and off at the zero crossing points of the mains cycle. If a frame period of 2 seconds is chosen then by deciding how many cycles are ON during this period, the power can be adjusted in 1% intervals (assuming a 50 Hz supply).

Note that full cycles are chosen in order to avoid giving a DC component. The designer is still constrained by EN61000-3-3 to prevent the lights flickering due to pulsing load currents.

18.9 Radiated heat transfer

Radiated heat transfer is a subject that is often kept in its own little box, never to be disturbed by actual use. You might remember from school that radiated heat transfer follows the forth power of absolute temperature, but this fact has never been put to use by you or any of your colleagues. Let's investigate its relevance to small-signal electronics.

Firstly, that half remembered equation you were thinking about is the *Stefan-Boltzmann Law*; namely: $P = \sigma\, A\, T^4$, where the Stefan-Boltzmann constant, σ, has a value of 5.67 pW/(cm²·°K^4). Let's look at a 1206 surface mount resistor; 250 mW power rating. It has a radiating surface of 0.12×0.06 inch^2, which is 0.046 cm². At 150°C the maximum possible radiated emission from this component is 5.67×0.046×(273+150)^4 pW= 8.4mW. Now this assumes that the resistor is a perfect emitter [A Black-Body, which has an ***emissivity*** of 1 by definition] and that it is radiating into cold space at 0°K. The point is that for small surface-mount components at ordinary temperatures you can just neglect radiant heat transfer completely.

Let's try another example. Consider a hot plain heatsink with a radiating surface 10 cm × 5 cm. It is at 100°C in a 20°C ambient. The maximum possible radiant heat transfer is 50×5.67×[(273+100)^4–(273+20)^4] pW=3.4 W. Notice that this time I have taken the ambient temperature into account, but I am still assuming that the heatsink has an emissivity of 1. The radiated heat transfer now looks pretty significant. Let's look at the lowest possible convected heat transfer for comparison. If the plate were to be a horizontal surface facing down then you would get the minimal natural convection.

$$\boxed{\textit{heat transfer rate} = h \times A \times \Delta T}$$

$h = 0.58 \times \left[\dfrac{\Delta T}{L}\right]^{0.25}$ which gives 0.58×[80/0.1]$^{0.25}$ ×0.1×0.05×80= 1.2 W. Hence the radiated cooling effect is important in this case.

The emission spectrum of a hot (ideal) body has a pronounced peak at a wavelength given by *Wien's displacement law*: $\lambda_{PEAK} = \dfrac{2900}{T_{abs}}\,\mu\text{m}$.

Temp	λ_{PEAK}
20°C	9.9 µm
50°C	9.0 µm
100°C	7.8 µm
150°C	6.9 µm
200°C	6.1 µm
250°C	5.5 µm
500°C	3.7 µm
1000°C	2.3 µm

Whilst the peak of the response is still in the infra-red, the visible content increases so that red heat is seen on an iron rod at around 700°C. The peak response does not fit into the middle of the visible spectrum until around 6000°C.

Wavelength	Approximate Colour
100-15 µm	Extreme infra-red
15-6 µm	Far infra-red
6-3 µm	Middle infra-red
3000-750 nm	Near infra-red
750-630 nm	Red
630-600 nm	Orange
600-580 nm	Yellow
580-530 nm	Green
530-470 nm	Cyan
470-440 nm	Blue
440-420 nm	Indigo
420-370 nm	Violet
370-300 nm	Near ultra-violet
300-200 nm	Far ultra-violet
200-10 nm	Extreme ultra-violet

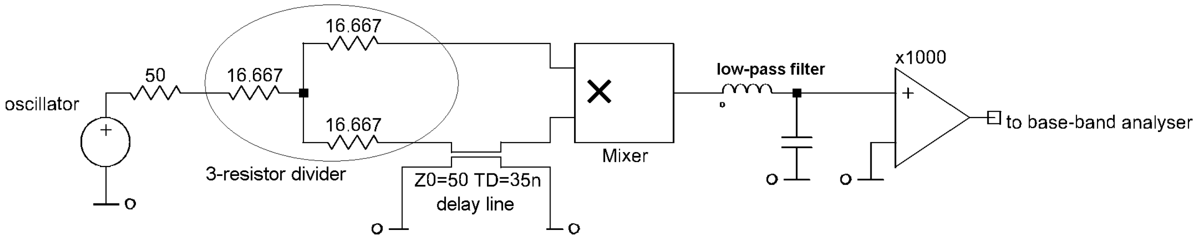
Measuring Oscillator Stability/Phase Noise

Mix the oscillator with a delayed version of itself using a long transmission line;[5] this is an *auto-correlation* technique.

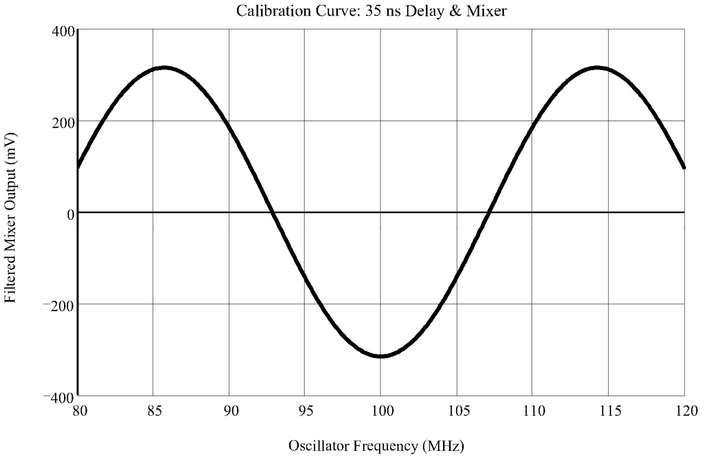
Taking the mixer as an ideal multiplier, the result of the mixing process is:

$$\sin(2\pi ft)\times\sin(2\pi f[t+t_d]) = \frac{1}{2}[\cos(2\pi ft_d) - \cos(4\pi ft + 2\pi ft_d)]$$

The double-frequency term is filtered by the low-pass filter, leaving an output voltage dependant on the phase shift produced by the delay line. The sensitivity of this scheme is poor. 18 dBm input through the 6 dB splitter gives 12 dBm at the mixer LO input. The delay line should be lossy or attenuated to give 6 dBm at the mixer RF input. If the mixer *conversion loss* is 6 dB, this gives 0 dBm output (630 mV ptp).



Using a variable frequency oscillator, the calibration curve for this *delay line discriminator* system is plotted without the 60 dB LF amplifier present. (The amplifier would otherwise limit {clip})



There are two aspects to the gain: the ptp voltage, and the frequency difference between peaks. The delay line length is adjusted to give 0V mixer output at the frequency of interest. This gives the largest gain, but also gives optimum linearity.

Small frequency fluctuations produce proportionate voltage fluctuations.

From the curve, the transfer function slope is seen to be approximately 9.6 mV/MHz in this case. (If the delay line length was doubled, the transfer function slope would also be doubled.) Using the 60 dB low noise amplifier then takes the slope to 9.6 V/MHz or 9.6 µV/Hz. The 'base-band analyser' could be a DSO with FFT capability, since it does not require a bandwidth greater than 1 MHz.

This delay line discriminator method requires **very** careful interconnection in order to minimise noise. The 60 dB base-band { LF } amplifier should ideally have differential inputs in order to minimise noise caused by ground loops.

[5] R.T. Adair, R.L. Ehret, and E.M. Livingston, 'Measurement of RF Signal Generator Phase Noise Using a One-Generator Delay-Line Method', in *IEEE Transactions on Instrumentation and Measurement*, IM-35, no. 4 (Dec 1986), pp. 496-502.

CH19: lab/workshop practice

19.1 Dangerous Things

There are many practices that are strictly based on the history at any particular company. These are not necessarily without merit and in any case you will need to comply with the *local traditions*. However, there may come a time when you will be in a position to change these traditional values, and you should therefore make up your own mind as to their worth.

Any advice given in this section should be checked against the working standards and practices in the industry and location where you are working. This also applies to the various discussions on safety and tolerancing. If my recommendations are more stringent then I advise you to follow them. If your company requirements are more stringent then you must follow those, *or get them changed.* To do otherwise is to get yourself into trouble.

Safety is a key requirement for any workshop or laboratory environment. It is unacceptable to get injured yourself, or to allow injury to occur to others in the area. For electric shock hazards, 25 V AC and 60 V DC are considered safe. Above that, depending on the area of skin contact, and the presence of moisture, electric shock becomes a possibility. However, electric shock is not the only hazard from electricity. Other electrical hazards include arcs, burns and explosions. These can occur where the product of short-circuit fault current and nominal working voltage reach too high a level. Thus automotive 12 V batteries do not present a shock hazard, but the fault current is sufficiently high to produce dangerous arcing. Thus even 12 V supplies should not automatically be considered as “safe”.

If a large capacitor is charged up it can provide a source of high current and hence an arc hazard. Remember that the energy stored in a capacitor is $\frac{1}{2}CV^2$. Electrolytic capacitors are a special case for size considerations. For a given technology, the volume is proportional to both the capacitance and the working voltage.

EX 19.1.1: Capacitor A is rated at 30 V and 10,000 μF. Capacitor B is rated at 300 V and 1000 μF. The capacitors are from the same family, and use the same technology.

a) How do the sizes compare?
b) How do the maximum stored energies compare? (Answer on p.399)

Modern electrolytic capacitors have vents in them so they do not explode as violently as they used to (prior to say 1990). However, given that they can still eject fluid and material at high speed, you would do well to avoid being in the ‘line of fire’ when a PCB containing electrolytics is first switched on. An electrolytic that is placed in circuit backwards {reversed polarity} will draw a lot of current. If the power supply is capable of supplying this current one of two things can happen: One, the capacitor gets hot very quickly and explodes, or two, the capacitor gets hot, but not hot enough to show visible damage. The capacitor may just sit there overheating for hours, days, or weeks before it becomes noticeably defective. The biggest bang will occur with a capacitor which behaves as a good short-circuit when connected across a power supply with a high power output capability.

I usually put a transparent plastic screen over a new board when it is first switched on, or when I am probing around a high power board. Eyes are easily damaged and equally easily protected by a simple screen. You can work on these circuits for twenty years and not have a problem. Then one day, when you are tired or distracted, your probe might slip, causing you to short-out a power rail; you can then end up regretting the mistake forever afterwards. An alternative is to just wear safety glasses, although you do still risk getting exploding things on the rest of your face.

You also have to be aware that fuses can explode violently, especially glass bodied fuses. A fuse has to be rated to break the maximum possible fault current. If this rating is exceeded then the fuse may physically explode. *Low breaking capacity* fuses are usually glass bodied and these are ideal for spraying glass fragments around a development area. Even fuses hidden away in fuse holders can burst forth from their enclosures with wild abandon. If there is even the slightest possibility of there being too large a fault current, a *high breaking capacity* fuse must be specified. Fuses have their own section in the next chapter.

19.2 Soldering

Soldering, as a means of joining metals together, has been in existence since around 2500BC and it is likely to remain in use for the foreseeable future. The *Reduction of Hazardous Substances Regulations* (2006) have largely removed lead from solder formulations for environmental protection, but the concepts remain similar.

The ability to solder well is a vital physical skill for an engineer to learn. It is unacceptable to leave this as the sole responsibility of the lab technician. To do so is an admission of the lack of the necessary skill. If you have never been shown how to solder, it is understandable that you may be unskilled in the process, and perhaps too embarrassed to admit it. This will be a failure in your training, actually, rather than your fault. However, when you recognise that your training has been inadequate, it is negligent of you to not take steps to remedy it. If you have to, watch the lab technicians and/or get them to show you how to make *good* joints. The process is relatively simple. You can be shown the basics of how to solder in a few minutes.

There was a software engineer where I was working who had done a computer science course to masters level. He was working on embedded processor systems and yet couldn't read the circuit diagrams or handle a soldering iron. In our environment he was like a cripple. He just needed a bit of a push to learn how to solder and a bit of help on a few basic circuits and he was much more effective at his real job. The fact of the matter is that even software engineers occasionally need to attach test points to the circuit to see why their code is not driving the hardware correctly.

When I started out as a trainee, I was working under an engineer who would look over the prototype I had made and mark unsatisfactory solder joints with a red felt tipped pen and get me to remake them! I can't say that any of them were ***dry joints***, but his view was that the joints had to look 100% perfect. He did not want to spend time debugging a prototype because of a faulty joint.

To make a good soldered joint the two conducting surfaces must be clean. New components on freshly made boards are automatically clean. Old components and old boards may not be. Do not be afraid to use a knife, the end of a flat bladed screw driver, or even a pair of side cutters, to gently scrape an oxide layer off the leads of an old oxidised wire-ended component. Sometimes I just lightly grip the lead in a pair of pliers and pull the component body using my fingers. The plier jaws sliding down the lead take the oxide off that part of the lead. You have to repeat this process a few times, rotating the component on its axis each time, to get a clean lead. This naturally only applies to old wire-ended parts. For oxidised surface mount parts, the only thing you can do is to get some extra liquid flux to apply to the joint prior to soldering.

Whilst the aggressive fluxes used prior to say 1990 were excellent at removing oxides, modern low-residue, 'no-clean' fluxes are less aggressive and less effective. This means you have to be more careful with joint preparation than used to be the case.

You are supposed to apply the solder to the joint and not the iron. This stops the flux burning off before it has a chance to do its work on the surfaces. But don't get too literal about this. The iron needs a slight film of solder on it as well. This makes a better contact to the joint. The key is speed; the joint needs to be made quickly. If you have to hold the iron onto the joint for more than a few seconds before the solder melts [let's say 5 seconds] then the iron is not big enough [not enough power] for the job. In fact if you are soldering tin-plate or brass screens to your circuit board you may well need a larger iron then the lightweight types used for surface mount components. I like to have one large iron and one small iron on my bench for this reason.

Sometimes you need to use two irons when soldering heavy brackets and the like where three or more hands would be useful. Ok. Press gang {conscript; force to 'volunteer'} anyone passing by to give you a hand.

If any joint you make does not look good then don't be afraid to remake it. Suck away the old solder with a solder sucker, or with desoldering braid, and remake the joint. This is not an admission of failure to solder. In fact, leaving a poor looking joint is the admission of failure. If the joint was a bit dirty it might have needed a bit more flux than was available; remaking it will give the joint that extra flux. Whilst modern 'no-clean' fluxes leave minimal residues, being designed to not require cleaning off after the soldering operation, you should be aware that these fluxes are not good at cleaning oxides off of joints! You are therefore likely to get a poor joint if the conductors are even slightly oxidised.

Solder joints from wires to terminal pins are best made by wrapping the wire around the terminal prior to soldering. In fact this is required for earth/ground wires; the earth/ground wire must not be mechanically fixed by solder alone. This ***wrapped joint*** is much better at resisting general vibration and movement of the wire and is therefore a stronger joint for any soldered interconnection. The wire is also held rigidly during the soldering operation, thereby removing one reason for a poor joint. The only problem with a wrapped joint is when you come to undo it.

The natural thing to do is to heat the joint and pull on the insulating sheath of the wire. The problem is that solder specks then tend to flick about. I have been very concerned to find flakes of solder on my glasses. Without the glasses, the solder and flux would have been in my eyes! When making test boards which are going to be dismantled very soon after being built, I therefore tend to use a plain butt-joint on the interconnects to make them easier to undo. Otherwise I would recommend cutting the wire near to the joint and pulling the strands out of the joint with a pair of pliers. Nevertheless, this wire removal process is probably the most hazardous soldering type activity that you will encounter, safety glasses being essential.

FR4 is standard epoxy glass reinforced PCB material. FR4 is what you normally get when you order a PCB from a vendor. However, not all FR4 is the same. For simple applications any FR4 material will do fine, but more demanding applications may require you to pick a particular type of FR4. One consideration is the operating temperature. FR4 has what is known as a *glass transition temperature*, T_g. At this temperature the material changes from being rigid to being soft and rubbery.

Glass transition temperatures vary from manufacturer to manufacturer, but will probably be in the range of 125°C to 140°C for standard material. High temperature material may extend this up to 180°C. If you plan to run the PCB above 100°C you need to start making enquiries of the PCB vendor. Some material is only UL rated to 105°C for example.

Notice that if a small area of the PCB is over-heated the board will not lose its overall structural rigidity. You may then be concerned about localised discolouration of the board and ensuring that there are no mechanical mounting points in this softened area. Also realise that some resistors are specified designed to run at body temperatures up to at least 155°C. If large surface mount resistors were to be run at their full ratings, then subjected to physical shock, it is possible that the adhesive used on the PCB would not be up to the job. Thus softening of the PCB material and the adhesive are critical factors when considering the operating temperature of an assembly that is subject to vibration or mechanical shock.

Dielectric constant is fairly consistent between different types of FR4; K values between 4 and 6 are to be expected. Other factors like moisture absorption, dissipation factor and dielectric absorption can also vary significantly between manufacturers. Correct board layout will minimise the intrusion of these imperfections in the board material, and therefore the differences between the types of material will be less significant.

19.3 Prototypes

Matrix board with copper strips, or other copper patterns, is a good way of supporting components for test circuits larger than 5 or 6 components. There are a few things that I would like to say about such prototypes. Firstly, you have to be able to make them yourself. If you have to rely on a laboratory technician to build your design because you are not able to do it yourself [by reason of skill] then you should acquire that skill. This is not to say that you should always build your own prototypes, but you should certainly be able to build your own prototypes because you will definitely want to be able to modify them yourself at the very least. [Circuits very rarely work completely on the first try.]

I have seen engineers and technicians breaking off pieces of matrix strip board using a variety of 'ingenious' techniques such as breaking the board over the corner of a bench or holding it in a vice and snapping it. These methods are acceptable only if you heavily score {cut} the copper tracks first. If you do not score the tracks then you inevitably stretch the tracks and what can then happen is that you create a small fracture {break} in one or more of the copper tracks, making an intermittent connection. This is very unproductive, and if you had ever spent a day trying to hunt down an elusive intermittent fault caused by an over-stressed cracked track then you would understand what I am talking about and would never risk creating such a fault.

Whilst I agree that heavily scoring the tracks and snapping the board is an acceptable procedure, nowadays I just use a hacksaw instead. It really doesn't take very long and is a much more guaranteed process.

The next thing you have to do when making prototypes on matrix strip board is to make links between various bits of the circuitry. The easiest way is to use a large piece of board with, for example, the tracks running horizontally and the links running vertically. It makes a neat job and it is very easy to follow the circuit. I can't say that I ever plan the layout though. I just guess where it is going to go and start wiring. Planning it out is a bit over-enthusiastic and if the circuit is that large then I would either get a lab technician to make it or have it made on a PCB.

I used to use tinned copper wire and PTFE sleeving, but that is quite slow. I would now use plain 24 SWG tinned copper wire without sleeving. This produces a neat job if you know one little SEEKret. Tinned copper wire comes on small reels. When you unreel it, the result is a springy coily mess that goes everywhere but where you want it to go. When you make links out of this stuff they are not straight and tight like they ought to be. The wire needs to be calmed down; *passivated*. Take a good length of wire, say 1 metre long, and clamp one end in a bench vice. Hold the other end wrapped twice around the wide part of the jaws of a pair of pliers. Take up the slack then gently tension the wire. Now give it a pull until you feel the wire 'give' {stretch}. I could measure this and tell you the required elongation, but that would be very artificial. You just pull it until it starts to flow, then stop. Cut off the damaged parts at both ends of the wire and you will have a straight passivated piece of wire that will go just exactly where you want it to.

In fact this passivation technique has wider application. When you want to make a twisted pair, a very common task, get your two pieces of wire and clamp the two wires together at one end in a bench vice. Pull the wires together along their whole length until you reach the other end, then trim the ends to the same length before inserting into the jaws of a hand drill or an electric drill. Operating the drill now produces a nice even twisting of the wire, much neater and faster than you would achieve by weaving the strands together by hand. Now there are two things you can do; either over-twist the wire and let it untwist a little when you release the drill chuck, or, when you have finished twisting, give the wire a pull to 'set it' in position. I usually just pull the drill body back until the wire slides out of the chuck. You now have a nicely twisted pair, but again you must cut off and discard the first and last few centimetres of the wire.

You have to apply a bit of common sense when using strip-board in particular. If you have nodes in the circuit which are sensitive to capacitance then put a break in the copper strip after the last connection points at both ends, thereby minimising the capacitance. If you are working with high voltages, say >100 V, you will need a greater ***creepage*** distance; just peel back unnecessary copper strip around that area. Use a scalpel blade or a sharp knife to lift up the copper whilst applying heat with a soldering iron. This is a useful skill to acquire, and it is best if you get somebody to show you how.

It is not the best designs that are produced, but those that are *perceived* to be the best by those who make the decisions.

If your prototype unit looks like a pile of junk, it may well be treated with less respect than if it had looked a bit better. This may not be sensible, or clever, or popular; but then you have to deal with managers who may not be the most brilliant technical beings in the universe. This being the case, it is as well to make the prototype look reasonably tidy.

Dressing the wires nicely, or filing the edges of the matrix board, may seem unimportant to you, but if those things sway the decision makers then they should be done. In order to be a 'success' you have to get your *good* ideas produced or enacted {made to happen}. The more of these successful designs that are made, the more of a success you are; better salary and job prospects result. A 'designer' who has no designs in production for an extended period is no use to anybody.

If anybody tells you that you can't make a prototype on matrix board then you should get very suspicious. I accept that the performance will never be quite as good as a well done PCB, but I have produced lots of quick prototypes on matrix board to test out an idea before going to a PCB and this has saved lots of re-iteration of the PCB and therefore saved both time and money. (mm-wave and microwave circuits are excluded from this prototyping argument. They typically have to be prototyped on special ceramic substrates.)

I made the first prototype of a 200 ps edge speed, 20 MHz repetition-rate pulse generator on matrix board before going to a PCB and that was a useful thing to do. The response was not quite as good as the PCB (perhaps 250 ps with a bit more ringing), but nevertheless demonstrated the workability of the circuit. At the other end of the spectrum, I built a 0.1% accuracy flat pulse generator at 1 kHz repetition-rate on matrix board. You can build sub-pA bias current amplifiers and sub-microvolt amplifiers on matrix board as well without much difficulty, if you know what you are doing. Just remember to peel off the copper where it would be a liability. If necessary, paint guard tracks on the board using conductive ink (eg silver loaded ink). You can even suspend a few critical joints in the air above the board. This is a good discipline because you have to think about which tracks are critical. This information can then be passed on to the PCB layout person (or remembered by you when you come to do the layout yourself.).

You must be able to operate the PCB layout package that your company uses at least to a limited extent. It is vital to closely control key areas of a layout and it can be too frustrating for the PCB layout people to have you sit there whilst you try to tell them which wire goes where. It is often easier to just do the layout for those critical sections yourself, perhaps with the PCB layout person on hand to assist.

Actually it can be very difficult dealing with PCB layout people when you are new to a company. You have all sorts of ideas of how things ought to be done and they quote all sorts of rules at you. A good one is 'copper balance'. PCB manufacturers like to have equal amounts of copper on all layers or they will sometime allow balance between adjacent pairs of layers. This is important for surface mounted assemblies; an unbalanced board will tend to warp {twist} when heated. [For example when it goes through an infra-red reflow oven or even when it is made.]

Better processing means that the PCB is held in a solid press when it is cured and it is clamped when it is reflowed. The technology is there to make that problem a non-problem. The problems you get into by following the PCB designer's advice are worse though. If there is an area of board that must not have ground and power planes in order to minimise the capacitance, then that is what you *have* to do.

Another ‘trick’ PCB designers come up with is a cross-hatched ground plane. This again reduces the mechanical stress on the board. I allowed this once, and the board was a disaster, even though it was only running at modest frequencies [probably 200 MHz clock frequencies and 150 MHz analogue bandwidth]. I would never want a cross-hatched ground plane on any future design as a result. They just don’t work due to the increased inductance.

Multilayer boards have pads on all layers as a rule, even if there is no connection to that pad. I normally specify that unused pads on inner layers are removed as a precaution against excess capacitance and leakage. It has to be said the these unused pads will hold the plated-through hole metallisation in place better if a through-hole component is being desoldered from the board. However, I don’t use too many through-hole {wire-ended} components nowadays, so I am not bothered by that. Also modern via {stitch-thru} hole plated technology is so good that the plated through barrel is very hard to remove even when you deliberately want too.

On the subject of vias, don’t try to save on the number of vias used by taking several tracks to the same ground plane via. It is better if each component uses its own via. Otherwise the track and via impedance will allow RF voltages to couple from one of the components to the others, thereby creating noise problems, and even oscillations.

Another thing to be aware of is *thermal reliefs*. If you track a component into a ground plane, most PCB packages can be configured to provide thin interconnects to the ground plane from the component pad. This makes it easier to unsolder the component because you don’t have to heat up the whole ground plane to de-solder the component. There is no problem ***wave soldering*** the components because the solder wave heats the whole board up in that area, but de-soldering with a soldering iron can be difficult. I accept that it is more difficult and I just tell them to use a bigger iron! This is a good way of making yourself unpopular, but if you need to put high currents through the tracks then a few 0.008 inch tracks is not going to be adequate. Also, if you want minimum inductance, the last thing you need is a thin track.

The PCB layout is critical to any sophisticated design. It is important on high voltage circuitry [>100 V], on low current circuits [<1 nA resolution], low voltage circuits [< 100 μV resolution], high current circuits [>500 mA], and circuits operating above a few megahertz. I find that there are always parts of each circuit board that need special consideration. Only the most trivial of circuits can be left entirely to a PCB person without supervision.

19.4 Experimental Development

There is a very big difference between a development lab and a calibration lab. A development lab should have racks or tubes of components. It should have facilities for making modifications to prototypes such as a workbench, bench vice, hand drill, files &c. These things are best kept completely separate from a calibration lab. In fact no circuit work of any kind should be done in a calibration lab. It should be a ‘clean’ environment without so much as a soldering iron. It is not uncommon for both facilities to be provided within one company. In this case they should ideally be kept in separate rooms with restricted access between them.

I worked for a short time in a development area where there were no components available for modifications. If you wished to change a resistor value you had to go to another building and get a component out of a locked cabinet! This company did defence related work, but this is no excuse for such gross inefficiency. It is necessary to find the correct value of a component to make a particular circuit work optimally. This can be done by calculation; it can be done by computer simulation; it can be done by using a resistance box and 'dialling up' a value until the circuit works then reading off the value; it can be done by adjusting a variable resistor then measuring its final resistance; or it can be done by changing component values on a PCB. All five are valid techniques in their appropriate place.

Don't let anyone tell you that only the calculation method must be used or you are not being professional. I have already explained about the characteristics of a professional. You have to get the job done. Do it the most efficient and effective way possible. Usually for simple circuits you would calculate the values and that would be the end of it. If you were dealing with a non-linear circuit for frequencies below say 1 MHz, then you might get away with using a resistance box to set a value in a circuit that was a bit difficult to characterise by calculation or simulation.

The underlying point here is that by using a variable resistor or changing components on the board you are effectively doing a 1:1 analog simulation of the system. What could be more accurate? The non-linear devices would not need to be specifically characterised, the answer would sort itself out. Now you may well find that you get the circuit working quickly this way then go back and do the tolerancing. I explained about this before as well. You want a working circuit to tolerance. If you can't get it to work then that is a good reason not to tolerance it.

I have found that the best circuits work when you build them with almost any values in place. When you put in the right values they work better. If the circuit is so critical that it won't work until all the values are spot on {precise; accurate} then you are going to have trouble with that circuit.

As you move up higher in frequency, your options become more limited. Using a resistance box in the main signal path of a >30 MHz circuit probably won't work at all. You will be better off using a small cermet preset pot. Above say 300 MHz even that method won't work, you will have to do direct substitution of components. I know that this is 'unscientific' and will be academically unpopular; I'm just telling you how it is.

If you want to see how best to match the impedance of a particular track then the best thing to do is to try it with a 47 Ω resistor and see if it needs to be higher or lower. It will probably be in the range of 33 Ω to 100 Ω, so there is not much of a range to have to try. There may be too many tracks to define as having controlled impedance, and when a track changes layers and goes through vias it becomes impossible to calculate the impedance.

If you have such components to change then you had better be very good at soldering. Somebody hand-building a board has to solder the component down once. You will have to solder and desolder the components many times. This means taking components off a PCB many times, and in a manner that is non-destructive to the PCB. Mostly you can just throw the old component away, but there will be times when you want to try the same IC or transistor again. In any case you must minimise damage to the PCB. You do this by applying heat for the minimum amount of time.

For expendable wire-ended components, cut the legs off close to the body then pull on the component leg with a pair of pliers after applying heat to the joint. Note: If the component leg has been bent over when it was soldered, it will be more destructive to pull the lead through from the component side. In this case straighten up the component lead on the component side of the board and pull it through from the other side.

Clean the hole with a 'solder sucker' of an appropriate size, holding the soldering iron on the opposite side of the PCB to the solder sucker if possible. [This is for plated-thru boards.] If you need to reuse the component then a good solder sucker can be used to free the component leads one at a time, but you stand a lot more chance of pulling the barrel of the plated through hole out of the board. Better still, use a specialist desoldering irons with built-in suction. These are particularly important when non-destructively removing components with more than three legs.

For surface mounted components there are different techniques. For two terminal components such as resistors and capacitors you can use two irons. You can then heat both joints simultaneously; this gives minimal overheating to the PCB and gets the component off of its pads in say one second. The use of one iron and a large blob of solder is not a good technique. Moving the iron between the ends of the component until the solder melts both ends at once does work, but again it does more damage to the board in terms of the number of times you can do it before the tracks peel off. Heated desoldering tweezers look as if they should also work well, but I have never tried them.

For SOT23 or smaller three-legged surface mount, devices it is easier to get a pointy tip iron and use the tip to lift up the single leg. Now you can take your pick of the other two legs. Four legged devices are more tricky and you have to start using desoldering braid {desoldering wick} to soak up the excess solder. Then you can use a sharp modelling knife to help lift the individual legs off of the board.

Now that manufacturers are using more leadless devices you also need a hot-air gun. The gull-wing leaded devices can always be removed with a modelling knife, but desoldering braid is not guaranteed to get rid of all the solder on a leadless package. Actually a hot air gun is also very useful on SO8 ICs, but if the packing density is very high, a hot air gun, even with a shaped nozzle, can be difficult to focus just on the one part you need to remove.

There was an old technique of *consolidating* stranded copper wires by dipping the ends in solder or tinning them manually with an ordinary soldering iron. This practice has been found to be unsatisfactory for long term use in screw terminal applications. The problem is due to an effect called *cold flow*. If you ever have occasion to pull apart mains plugs where the wires have been tinned, you will see the effect. The pressure on the solder causes it to creep with time and the connection becomes loose. This loose connection has a high resistance and can therefore overheat. This is the reason for the statement in safety standards that wires should not be tinned before they are crimped or screwed.[†] As a guide, a pressure of 1 MPa {1 N/mm^2} will cause (cold) solder to creep by 0.01%/day, this stress level being 100× lower than the tensile strength of solder.

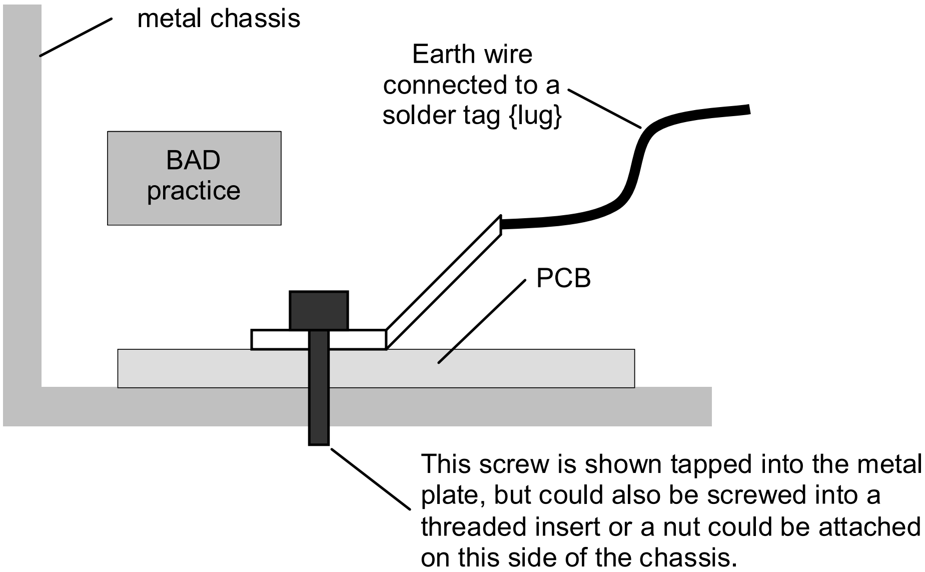
Creep is also expected in plastics and it is therefore unsound and unacceptable to have plastic involved in the pressure applied to a joint. For example, a ground {earth} bond that involves a screw connection clamping two metal faces together must not have an

[†] EN60950-1:2006, §3.3.8

insulator in the compression path.[‡] The plastic could creep and the joint would then become loose. The only exception would be if the compression were to be applied via a substantial multi-turn spring washer [not a single split-washer]. In this case it could be argued that the spring force of the washer would take up any creep in the plastic.

Don't make ground or power connections like this. As the PCB creeps under the pressure, the joint becomes loose and the connection is no longer guaranteed.

FIGURE 19.4A:


metal chassis
Earth wire connected to a solder tag {lug}
BAD practice
PCB
This screw is shown tapped into the metal plate, but could also be screwed into a threaded insert or a nut could be attached on this side of the chassis.

19.5 Screws and Mechanics

There are principles of mechanical design which are arguably more associated with automotive engineering that electronics. Screws which hold PCBs to the instrument chassis must not come loose in transit. This is prevented by use of shake-proof washers, self-locking nuts (eg Nyloc), or some sort of thread lock. Arguably this is somebody else's job, but if you don't think about such things, the project could fail, taking you down with it.

Remember that as you become more senior you will be expected to widen your horizons to encompass project management. Thus you will have to be considering 'boring' mechanics because everything has to be right including documentation, packaging, costings, delivery schedules and so forth.

How about making screws for a prototype the right length. Well you use a hacksaw on an over-length screw and cut it down. But how about screwing on a nut before cutting the thread? Once the thread has been cut and filed, it is easier to make the end good by unscrewing the nut over the cut portion of thread. Even graduate mechanical designers don't necessarily know this because it doesn't get written in text books!

What about an assembly held on by 10 screws. Do you put the screws in one at a time and tighten them? Only if you have never worked on a car! You always put all the screws in loosely before you tighten any. This action jigs the part into the right position.

[‡] EN60950-1:2006. §3.1.7

Tightening any screw locks the assembly before it is suitably positioned. This is something any automotive engineer would know.

When wrapping a wire around a screw terminal post, does it matter which way round it is wrapped? Wrap the wire anti-clockwise and the action of tightening the screw throws the wire off the post. Wrap the wire clockwise and tightening the screw pulls the wire tighter into the post. This is something any electrician would know.

PCBs with multiple power rails are slow to debug if you have no test points or test points at random positions all over the board. If you neatly arrange a row of test points for the power rails at one point on the board it can reduce test time and therefore cost. You wouldn't want to route sensitive signal all to one common area, but power rails go all around a board anyway so it doesn't take much effort to centralise the test points. Any experienced PCB designer would know this.

On PCBs with hand fitted electrolytic capacitors, it is convenient to have all the positive ends pointed in the same direction. This makes a visual check on the PCB much faster, and means that it is less likely for an electrolytic to be put in backwards. Any test technician would know this.

Do you get the point? There are lots of little tricks of the trade that people will know that you don't. Some of these would have been taught on apprenticeships, but these were given a bad name by poor instructors and are rarely found in this fast-paced cost-conscious modern world we find ourselves in. Employers now expect somebody else to train their staff due to the high cost, the result being that less training is occurring.

You must keep your eyes and ears open, picking up these SEEKrets from even the most stupid of people. Try to evaluate the idea rather than the source of the idea, as even stupid people occasionally have good ideas!

****

ANS 19.1.1:

A) Capacitor B is 10× larger in voltage and 10× smaller in capacitance. You can therefore expect that it will be the *roughly* same size as capacitor A.

B) $\frac{1}{2}C\cdot V^2 \rightarrow \frac{1}{2}\left[\frac{C}{10}\right]\cdot[10\cdot V]^2 = \frac{10}{2}C\cdot V^2$ The stored energy goes up by a factor of 10× for the same case size, the higher voltage capacitor storing more energy. Therefore it is not safe to judge an electrolytic capacitor's destructive ability purely by its size. Also power factor corrected switched-mode supplies, which use an intermediate DC voltage, are smallest when the intermediate voltage is high.

CH20: electronics & the law

20.1 Regulations & Safety

Electronics has come of age. Whilst electronic devices have been important to people throughout the twentieth century, it is only since the early 1990's that the general population has been so inundated with electronic items. Now electric power, telephone and radio services have been considered essential for almost a century, but these were always 'self-contained' industries, with their own rules and procedures. Now that electronic devices have so deeply penetrated into people's lives the politicians have stepped in, giving much stronger safeguards to the public.

It has been the case for many decades that equipment must not interfere with radio and TV broadcasts. It has also been required that certain products meet electrical safety standards; now however, all electronic products are required to meet stringent emission, immunity and safety standards.

Instead of letting consumers decide on the standard of products by voting with their wallets, electronic equipment is required to perform to certain minimum standards before it can be legally sold. In Europe this whole legal situation has become fully developed since around 1996 with a whole raft of ***regulations*** and associated standards. Many of these standards, or similar standards, have been around for decades, but with the new legislation it has become mandatory {compulsory} that at least some of these standards be met. The penalties are very explicit. If you make and sell equipment which is not safe according to the standards, these standards representing "good engineering practice", then that equipment can be banned from sale throughout all of the European Community. Because of mutual recognition agreements, standards met in one economic area are often acceptable in another area. Thus most of the world is now covered by this type of rigorous legislation.

Additionally there are penalties for those individuals who are guilty of producing this equipment. They can be held *personally* liable for their conduct. An engineer who falsified a report in such a case could be sent to prison and/or fined.

This legislation is a relatively recent development, but should not be taken as cruel and unusual treatment of engineers. There have always been food inspectors, works safety inspectors, building inspectors and the like. Electronics, excluding military, medical and aviation electronics, has been an amateurish sort of activity as far as the law was concerned. Now it has come of age and is being treated as a more professional activity.

As things stand now, if you do electrical work on civil engineering projects such as bridges or office blocks, you have to be a *qualified* Engineer and have your name up on the site boards. The crazy thing is that you could currently be producing electrical muscle massagers with no qualifications at all. Somebody has to sign a certificate declaring that the product is safe, but this person does not need to be qualified; if the product is later found to be unsafe that person can be jailed and/or fined. This is perhaps a bit late after the faulty design of the massagers has electrocuted {killed by electricity} several users in different parts of the country.

Given the increase in legal suits against everybody for seemingly ridiculous things, you should expect that it will soon be required that those signing such declarations to be members of the appropriate professional body. For most electronics designers this would be the IET (formerly IEE), or the IEEE, or a local equivalent. In some parts of the world, registration as a *professional engineer* is already required.

Let's look at the law as it stands in Europe. There are local equivalents all over the world matching these requirements, so don't think you can escape them! In Europe, equipment is marked with a special symbol-form of the letters CE, the European Conformity mark. The mark is a legal declaration by the manufacturer that the product meets *all* regulations {laws} relating to that type of product. If the product is manufactured outside the Community, then the importer of the equipment is accepting responsibility and liability for the product. In the USA there are UL and CSA standards to meet. In Australia they are called C-tick. Conformity marks with similar meanings exist throughout the rest of the world.

Safety standards are the ones that most people latch onto straight away. For the specific type of product that you are making there will probably be some sort of safety standard that covers it. It is up to you to decide which is the appropriate standard. There are particular standards for things like medical equipment, laboratory test equipment, hand-held probes, machinery, equipment for use in mines &c.

Reading through these standards, they have common themes {areas of concern}.

- ❑ The equipment must not be able to harm the user by electrical shock even if one component fails (single-fault condition).
- ❑ The equipment must not be able to catch fire and spread this fire to its surroundings.
- ❑ The equipment must not emit toxic amounts of smoke, even if its insides are on fire.
- ❑ The equipment must not be able to fall over onto somebody and must not in any case crush, rip, scrape or gouge their body.
- ❑ The equipment must not be so hot on the outside as to damage skin [unless it is a special heat-sink or heating area].
- ❑ The equipment must not cause injury during normal or single-fault conditions by means of sound, heat, light, pressure, radiation, electricity, gaseous emission, or any other way not explicitly mentioned.

These are all things that you should expect in a safety standard. The user should be safe, even when something goes wrong with the equipment. But some of the standards go farther than this. The instrumentation standard requires that safety related information be given in particular symbols. Voltage has to be represented by an uppercase V not a lower case v. The number and units must be separated by a space so "230V" is not acceptable, neither is "230 v", but "230 V" is. Nothing must be appended to the units, so "230 VRMS" is not acceptable either. There is a lot of fine detail and the only way to get it right is to read the standards carefully, making a checklist of things that need to be done to verify the product against the standard. To do less would be to make yourself liable to both a civil law suit and a criminal charge of *negligence*.

One area of great concern in the standards is the handling of the ***mains*** ground connection {earth; protective conductor}. There are specific requirements for testing the ground path at high current. It is not unusual to test the ground path at 25 A RMS for 2 minutes and to ensure that the impedance is less than 0.1 Ω. Additionally there is a rule stating that the ground wire must not be held on with solder alone. (If soldered, the joint has to be first wrapped around the terminal post for mechanical support). Note that solder itself has a resistivity some 10× higher than copper.

Any screw connection of the ground has to be vibration resistant so it doesn't come loose. Any pressure contact of the ground path must not be transmitted through a plastic material unless there is enough spring force present to allow for any possible shrinkage in the plastic. All of these rules have come from practical experience and have been encapsulated in the standards as "best practice".[1] With that status, you really have no option but to follow them.

When dealing with explosive atmospheres, electronic systems have to be double fault protected. In other words two components are allowed to fail open or short circuit and the system still has to be safe. A circuit called a *zener barrier* can be used to cross from a standard circuit region to the protected region. A pair of zener diodes is used so that one can fail open-circuit without compromising safety. A fuse would be used into the zener barrier. If the standard circuit should become live due to a power supply fault, the fuse blows before excessive energy can be delivered to the potentially explosive region.

20.2 Fuses

Fuses are another major target for safety inspections, particularly the main power fuse inside instruments. I am going to discuss this in terms of fuses inside instruments, but the concepts are common to all fuses. This is an area that changes from country to country and you have to check on the latest rules and regulations. For the USA and Canada, fuses are rated to UL and CSA standards at 115 V. They are rated to break a current of 10,000 A and are usually made of glass. The fuse rating is *that current which will blow the fuse*. For European 230 V fuses you have two types, ordinary and High Breaking Capacity (HBC). In this context *capacity* means capability.

European fuses are rated at the maximum current *which they can carry continuously*. They are specified to IEC60227. This gives an inconsistent situation, because if you have to fit a 230 V HBC fuse for Europe, this is not acceptable in the USA or Canada because the breaking capacity is not high enough.

It is very much harder to break {interrupt; stop} a fault current at 230 V than it is at 115 V. Whilst a UL/CSA fuse is glass and can break a 10,000 A fault current, an IEC HBC fuse will only break 1000 A at 230 V. The IEC HBC fuse also has to have a ceramic body. Glass bodied fuses will explode *violently* if they try to break too large a fault current.

The thing that the designer has to look at when selecting a fuse is the *prospective fault current*. If there is a fault, how large can the current be? This would be evaluated by looking at the mains wiring in the building and the size and length of the mains lead. Often the answer is that you need an HBC fuse.

The older name for an HBC fuse is HRC, a *High Rupturing Capacity* fuse. The name change is not arbitrary, as the new name tells you more about what is happening. With

[1] K.O. Smith, and J. Madden, *Electrical Safety and the Law*, 4th edn (Blackwell Publishing, 2002).

the old name you realised that the fuse would safely rupture {fracture; break} up to a certain overload current. This is not the action you are looking for. Under fault circumstances, there is a short-circuit somewhere and you need to switch this current off {interrupt it} in some fashion. If you used a switch which was not rated for the current then the contacts might weld shut and the fault current would continue. The same can be said of the fuse. It has to safely *break* the fault current path. If the fuse explodes that is not safe. If it arcs over and the fault current continues to flow then that is not safe either. By using the appropriate voltage and fault current handling capability in the fuse, you have made the fuse safe. You have not necessarily protected the equipment.

For fuse selection, the first thing you have to worry about is the switch-on surge into the equipment. It is not unusual for 230 V equipment rated at say 230 W to take a sudden in-rush current in the region of 50 A to 150 A when the power is first applied. The mains switch and the fuse have to be able to withstand this current, time and time again. If the fuse is rated at 1 A and has to take a 100 A surge current without blowing then it needs to be rated for this. The fuse has a speed rating and you have to select one which can withstand the short switch-on surge pulse. This would ideally be read off a graph of the fuse characteristic. The fuse selected in this case would not be a *quick-blow* type, it would be a *time delay* [anti-surge] type.

For very short duration large current pulses the curves given may not be adequate or the current pulse may have a non-rectangular shape. It is then necessary to quantify the current pulse to select the correct type of fuse. The integral of the current squared over time is calculated and compared to the ***pre-arcing*** I^2t rating of the fuse. The I^2t of the pulse should typically be less than 50% of the pre-arc I^2t rating of the fuse in order to prevent unintended (nuisance) breaking of the fuse link. The I^2t value is often referred to as the *Joule Integral*.

$$\text{Joule Integral} \equiv \int_0^\infty [i(t)]^2 \cdot dt$$

Even "quick blow" fuses are not quick to blow if the fault current is not much above the rated current. Let me be more specific. The international standard for 5 mm×20 mm fuses is IEC60127. The following particular data comes from EN60127-2:1991. For a low breaking capacity quick acting fuse the opening time is a maximum of 30 minutes at 2.1× the rated current. Now that is a quick acting fuse! In order to get the fuse to operate in under 20 ms, the fault current has to exceed 10× the rated value. A particular manufacturer may well have an improved spec compared to that required by the standard; it is therefore up to the designer to choose suppliers carefully. Second sourcing such a fuse could be problematic if they both meet the standard, but one is three times faster than the standard. Thus even for an uninteresting, old-fashioned, low technology component like a fuse, you have to be careful that another manufacturer's alternative is actually equivalent.

Let's consider a specific example of a 1 A fast fuse. The I^2t rating is 0.19 A^2s. If the start-up surge is considered rectangular and the peak is 50 A, will the fuse blow unnecessarily? If the pulse is 0.1 A^2s the fuse should be ok, so that means the maximum pulse duration should be less than 40 μs. If the pulse is longer than that, either increase the fuse current rating or use an anti-surge (time delay) type.

Just to make life interesting, there are two ways to rate {specify} a fuse. I mentioned this earlier implicitly but not explicitly; you might have missed it. IEC60127 rates a fuse at the current it can *carry* continuously. This is the European way of doing things. In the USA and Canada the rating is from UL198G and is the current that will *blow* the fuse. The values are different by about 20% for the same fuse. Thus you will find multi-national equipment rated like this: IEC (UL/CSA) 0.4 A (0.5 A).

The single-fault principle has a problem; if the single-fault does not cause an *obvious* immediate failure then it can go undetected for a considerable period of time. In this case it becomes increasingly likely that another single-fault will occur. In this situation, safety may be compromised. Let's take a simple example. If the ground {earth} wire breaks on a piece of equipment, it is not likely to cause an obvious fault in the equipment. It might be years later before another fault occurs; but if the insulation on the transformer fails and there is a fault path to the internal ground node of the instrument, there is going to be serious shock hazard. It is quite possible for this new fault to also go undetected, leaving the instrument case live.

There is only one way to deal with this multiple single-fault problem. The equipment has to have a maintenance schedule that checks for non-obvious single-fault failures that may have occurred since the last maintenance check. In the UK there are guidance notes which suggest that portable appliances are routinely checked for ground faults and insulation failures. This is known as **P**ortable **A**ppliance **T**esting, or PAT for short. These tests may be sufficient to reduce the probability of cumulative single-fault failures. Whatever the equipment or installation involved, it should be part of the safety checking procedure to see if specific single-fault failures can remain undetected and to put checks for these in the maintenance procedure.

You may wonder why it is unusual nowadays to have a fixed mains lead {power cord} on a piece of equipment. This is the safety standards at work. The "pull" tests on the mains leads are so severe that you have to be able to practically swing the equipment around your head using the mains lead without damage. It is a lot simpler to put in an 'IEC inlet'. [This is an industry standard power connection scheme for equipment; the standard being IEC60320.] The cable pull test is therefore not applicable.

Another problem that this connector solution handles is that of the *disconnection device*. Safety standards require a device which isolates the equipment from the mains supply. It is easy to get switches which are rated to CSA/UL and other standards at a specific current level, but when you look at the prospective fault current there is a problem. If you specify a high breaking capacity 230 V fuse then this will break a fault current in excess of 1000 A. What happens to the switch under these circumstances? The answer is that the switch is no longer guaranteed to be working. What you might expect is that the fault current will fuse {melt} the contacts closed. Thus if you have to change the fuse you also have to also check the switch. It is easier to say that the IEC inlet is the disconnecting device for safety purposes and that the switch is only a functional device, rather than a safety disconnecting device.

It may also happen that a fuse fails for no apparent reason. This is known as *fuse fatigue* and is due to the fuse heating up and cooling down each time it is used for any period of time. The only remedy is to replace the fuse to see if it immediately blows again. If it does blow again immediately then the real reason for the fuse blowing will have to be found.

20.2 EMC

Another major area of legal and technical concern is EMC [Electro-Magnetic Compatibility]. There has been concern about EMC for military and avionics applications for many decades; there have also been regulations concerning radio frequency emissions from equipment and installations. EMC is a much more stringent requirement though. The basis of EMC is that electronic equipment must coexist with other equipment in the same area without either interfering with the other equipment or being interfered with.

There is a world of difference between the problems associated with EMC depending on the nature of the equipment. Let me give you some examples of incompatibilities that have prompted the legislation.

- ☹ A petrol pump reading incorrectly when a mobile phone was used nearby.
- ☹ A fire alarm being triggered by a mains transient {spike}.
- ☹ An aeroplane navigation system disabled by laptop computer emissions.
- ☹ A faulty RF garage door opener [receiver] jamming police transmissions.
- ☹ An electrically controlled robot arm going haywire {smashing into things}.
- ☹ An intravenous drug delivery system changing the dosage because somebody uses a piezoelectric cigarette lighter nearby.

Each of the specialist applications will have its own specific standard, these standards continually being updated to provide higher levels of protection.

The standards are very “asymmetric” in terms of emissions versus immunity. Whilst the immunity requirement for radiated fields would be of the order of magnitude of 3 V/m, the allowed emissions would be more like 30 dBμV/m [0 dBμV is 1 μV, so 30 dBμV/m is 31.6 μV/m]. These requirements are so stringent {tough} that only the simplest types of equipment do not need to be shielded.

Other EMC requirements include immunity from static discharges, immunity from mains transients [spikes and power dips], limited RF emissions down the mains wiring [conducted emissions; the mains wiring acts as an RF antenna], limited harmonic currents drawn from the mains [helps the power supply company], limited current pulsations drawn from the mains [stops the lights flickering]. Testing equipment for compliance with these standards is both expensive and time consuming; one or two days evaluation at a specialist test house is now quite a usual requirement.

Quite clearly EMC can relate to safety. It does not take much imagination to see that automated intravenous drug dispensers can cause serious consequences when they deliver the wrong dosage due to EMC. Sometimes the specific safety standards therefore include their own EMC requirements.

20.3 Radiated Emissions

In this area the standards are dictating the design of equipment and the direction of future development. The emission limits are very stringent and very specific in their measurement techniques. This allows the possibility of the standards being “worked around”. The most troublesome work-around is the so called *spread spectrum* approach. A single frequency emission occupies a very narrow part of the RF spectrum and therefore it stands a good chance of not interfering with other nearby equipment. Something like arc welding equipment, on the other hand, produces a very broad spectrum emission and is known to be very disruptive to radio communication.

In the spread spectrum technique the clock frequency of a digital system is deliberately modulated to reduce the amount of mean signal power in any given bandwidth. The problem is that everybody has receivers set up to step through the frequency band, measuring each point over a relatively narrow bandwidth. Spread spectrum transmissions then show a very much lower amplitude than a pure carrier of the same peak amplitude. Unfortunately the spread spectrum signal will undoubtedly have a greater interfering effect. It is also much more likely to interfere with other equipment because it is covering all of a much wider band of frequencies. Hopefully the standards will be changed to plug this 'loop hole' {an unforeseen gap in a law or rule which allows crafty people to avoid something}.

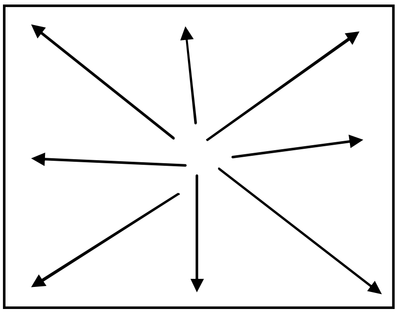
EMC text books like giving plane wave attenuation figures and formula. These are for the shielding effectiveness of slots, holes, seams and the like. These formulae are essentially meaningless for use on equipment and you should not use them at all. In fact the text books even admit that the formulae are for some academic situation that never exists in the real world, but they do not go on to tell you how to handle the situation.

No formula can give an accurate shielding effectiveness for a complex piece of equipment. There are major problems in predicting the performance of any design in terms of its electromagnetic radiation performance. Whilst simulation packages exist for predicting the emissions performance of electronic circuitry, the main problem is the complexity and repeatability of internal wiring. With defined wiring positions and defined assembly positions it is possible that a simulation model could predict the radiated emissions, but this would be a highly expensive and time consuming proposition. However, when you get different assemblies, with different options, different plug-in cards, different cable positions and so on, the computational task becomes impossible.

However, you still need to design things and you do need some sort of starting point. Do you need to puts screws every centimetre or can they be 20 cm apart? Do you need a mains inlet filter or is the design adequate without? There are definite rules for this sort of thing, but they are more empirical than theoretical.

Let's start with a fully sealed conducting enclosure. Let's say the box is made of metal and the sides are all welded together. There are no entrances or exits. Admittedly this represents an unrealistic "academic" situation, but one has to start somewhere.

FIGURE 20.3A:



Any electromagnetic radiation inside the box just bounces off the walls and cannot escape. The box does not radiate, nor is it susceptible to electromagnetic radiation. In order for this scheme to work the metal has to be a certain minimum thickness, this thickness being several times greater than the *skin depth* at the frequency being considered.

Now the metal is mostly reflecting the incident signal, so waves just bounce around inside. The field intensity inside the box will therefore be higher in places than if there were no box there at all. The safest rule to use, therefore, is that the ***shielding effectiveness*** is wholly due to the attenuation of the original signal strength.

The attenuation loss of a metal barrier is given as

$$\boxed{Attenuation = 8.7 \times \frac{\mathrm{t}}{\delta} \quad \mathrm{dB}}$$

t is the material thickness and δ is the skin depth.

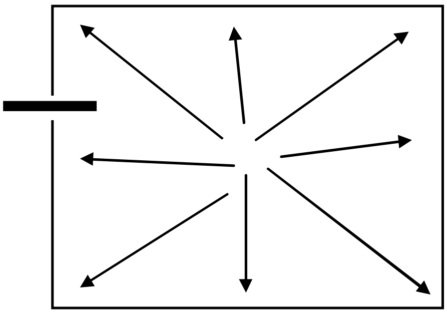
This attenuation formula comes from the approximation that the electromagnetic wave decays exponentially as it penetrates the metal, the decay being $\exp(-x/\delta)$ for a depth x. This decay, expressed in decibels is :

$$20 \times \log_{10}\left(\exp\left(-\frac{x}{\delta}\right)\right) = -\frac{x}{\delta} \times 20 \times \log_{10}(e) = -8.686 \times \frac{x}{\delta}.$$

The exponential decay itself is found by solving *Maxwell's equations* for an infinite conducting plane. This solution can be found in any good electromagnetics text book.

If you make the barrier at least 12× the skin depth [giving >100 dB theoretical attenuation] at the lowest frequency which may cause problems, you can then ignore the material completely. As far as the standards are concerned, only radiated emissions down to 30 MHz are of interest. If the case is made of aluminium then the material thickness is only required to be ≥0.2 mm. In practice, then, it is the construction of the box that is almost always the limiting factor.

FIGURE 20.3B:



If a small hole, say 2 mm in diameter, is made in the box, the shielding will probably not be severely compromised {reduced in effectiveness}. However, a simple piece of insulated wire inserted into the box through the hole will have a huge effect on the shielding effectiveness, even if the wire is not connected to any internal or external circuitry.

Now I have been using the technical term, *shielding effectiveness*. This is defined as the reduction in radiated emissions (in dB) as a result of a case {enclosure} or some other barrier. This is always the quantity you are interested in. If you could calculate the shielding effectiveness of an enclosure then you would have more control over the EMI characteristics of the equipment.

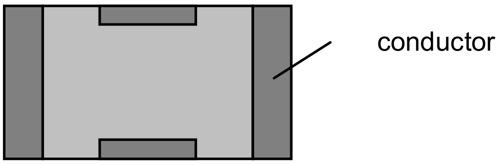
Whilst metal screened rooms can achieve shielding effectiveness figures of around 100 dB to 120 dB, you should not expect that sort of performance from a general piece of equipment with lots of I/O ports. In fact if you get better than 30 dB to 40 dB at all frequencies you will be doing well. It is actually very easy to achieve 60 dB at many frequencies, but what you find is that certain frequencies have considerably less attenuation of the internally generated frequencies.

I want to give you some figures to work with, so here is some data. To investigate the effect of an isolated piece of wire going through a barrier I put a battery powered 418 MHz transmitter inside an aluminium diecast box 220×145×100 mm. [418 MHz is a low power radio transmitter allowed frequency and these modules are readily available.] With an insulated 50 mm wire half in and half out of the box, the shielding effectiveness dropped to 54 dB. An insulated 200 mm wire dropped the shielding effectiveness to 35 dB. You should see that wires penetrating the barrier are a *major* problem.

Now this is not a new problem. In fact a special component has been developed for just this purpose; it is known as a feedthrough [or feedthru if you prefer]. The original style was a metal bodied part which was screwed to the chassis and which had solder tags on each end. Internally there would have been either a simple capacitor to the body or a more complicated filter consisting of inductors in series and capacitors to the body. With this sort of component the emission from the wire could be reduced by between 60 dB and 100 dB without any trouble. The difficulty is the mechanical construction.

This hand-wired assembly is a rather old-fashioned method which, whilst still workable and effective, is rather more expensive than is desirable. The key thing to achieve is to minimise the path length in series with the decoupling capacitors in the feedthru. This reduces the stray inductance and therefore improves the attenuation.

FIGURE 20.3C:



A modern alternative is a feedthru capacitor in surface mount form. The current flows easily along the length of the part through a low resistance. This should be <1 Ω. The centre conductors are the ground points of the capacitor. There are two in order to get a lower impedance. This part would either be mounted on a ground plane, or there would be via holes right next to the ground conductors. It is vital that these ground connections go straight to the outer enclosure chassis ground. Any length greater than say 2 mm would make the part relatively ineffective. Obviously there is not a sharp cut-off. More length makes the part progressively less useful in terms of its attenuation of RF energy.

Unless the equipment is powered by batteries and performs no external functions, it is clear that penetrating the RF enclosure is a necessary evil. The most likely essential is a mains {power} entry port. It is actually unusual to have a mains lead {power cord} directly passing through an RF enclosure on a modern piece of equipment. The usual technique is to use a chassis mounting power socket. This was discussed earlier, in the section on safety. This chassis mounting socket is an excellent opportunity to apply filtering to the mains lead.

Chassis mounting filtered inlets come in two distinct types: plastic body with a filter housing at the end and plastic body completely encased in metal. The type with the plastic body and the filter on the end are suitable for filtering emissions up to 30 MHz. That was their original function. They are only specified for this frequency range and their performance at higher frequencies is indeterminate {unknown}. I strongly recommend that you *never* use this type of filter for entry into a screened enclosure.

The point is that the filter body is not connected to the chassis and therefore it does not act as a feedthru at frequencies above a few hundred megahertz. If there is a screened enclosure, then presumably some shielding effectiveness is desirable. Going through the barrier with a plastic bodied filter housing will ruin the shielding performance of the barrier to the point where there is no point in having the barrier. Despite having a completely welded sealed box, use of this type of unshielded filter could reduce the shielding effectiveness down to a few decibels at certain (unspecified) frequencies. Don't do it!

For non-screened enclosures this type of filter may be suitable. It will give filtering at say <10 MHz and this may be adequate for your purposes. It will also be cheaper than a full metal-bodied part.

To be good at solving EMC related problems you have to be able to see the instrument or equipment from the point of view of an enclosed shell. Follow the surface of this shell around and see where there are gaps. With a full metal-bodied filter, the enclosure bends in around the filter and then pops out again. It is like having a dent in the solid shell and this does not cause a problem. I am assuming that the metal shell of the filter makes intimate {positive and complete} contact with the enclosure. This is **vital**.

I cannot overstress the importance of clean metal-to-metal [or conductor to conductor] contact for EMC shielding purposes. There is absolutely no point in having a metal case and then screwing it together with paint or other insulating finish on the mating surfaces. The mechanical department may whine {moan; complain} about the extra cost of masking the paint finish, but it is just something you have to do. Not only do you need clean metal faces, you need positive defined pressure on these surfaces, and you need surfaces which will not corrode. Untreated steel is useless from the corrosion point of view, even with indoor environmental conditions. Untreated aluminium and brass are not ideal for an indoor environment, but may be acceptable.

Perhaps you remember the chapter on relay contacts. If not then re-read that section. To get a good electrical contact you need *pressure*. Two pieces of metal just resting near each other will have some slight shielding effect; perhaps a few decibels. But if you screw, press, rivet or EMI-gasket the faces, then and only then, will you get performance of that joint up to the required level. Incidental contact {uncertain; not defined} is a nightmare when doing EMC work. You take the case apart and make a change; there is a difference. Was it the modification or experimental error? The only way to be sure is to pull the prototype apart and measure it without the modification. I have done this many a time, doing a sequence of modifications which should have brought the equipment back to the same state at the end, only to find that the equipment was now reading several decibels differently.

The experimental work associated with EMC work is very demanding. You sometimes need to repeat a modification a few times to ensure that it is actually doing what you think, especially when the modification is going to cost a significant amount of money. I have developed a procedure to deal with this and I strongly recommend it to you. First, when you get a new set of metalwork for the project you are working on, get one set made without paint on at all. This is not expensive. If you are having ten sets of metalwork made then the manufacturer can just leave one set of metalwork to the side when the others are sent to the paint shop. If the metalwork corrodes easily (steel for example) then you may like to get it plated with a good conductive finish such as electroless-nickel, or zinc with clear passivation. You probably won't think this advice is important, but you would have appreciated this advice prior to having to scrape off large areas of tough epoxy paint to check the effect of extra grounding points!

With EMC work you have to be prepared to get your hands dirty; and I mean that literally. To check for EMI improvements you need to block up holes, vents, seams and so on. They need to be free from paint and clean. This may involve wet & dry paper [or emery cloth, sand-paper, glass-paper] and elbow grease {physical effort}. It is also wise to rub the surface down with some sort of solvent to remove contamination, especially from sweaty fingers.

The SEEKret to making these tests is to use conductive adhesive copper tape [use an expensive high quality type] from the *outside* of the box. This gives a very repeatable and quick test. Are those vent holes causing the problem? Just tape over them and re-measure. The instrument will probably not overheat in the five minutes it should take to do the measurement. You can put the tape on and measure, take the tape off and measure, and repeat this a couple of times if necessary to convince yourself that it does or doesn't have an effect. Don't reuse the tape though; that is a false economy. Oh, and don't use the tape for production or as a long term solution. This tape is great for quick tests, but it is not great for long term use. Its joint resistance goes up by orders of magnitude when it is exposed to humidity and the long term performance is therefore very doubtful.

This use of copper tape will immediately tell you if the spot welds are too far apart or if you need more screws or if the gasket is not being compressed sufficiently. It is an excellent diagnostic method.

People are always trying to sell near field probes for this purpose. You are supposed to move the near field probe around the instrument looking for 'hot spots' of emission. I must say that I have not had much success with this technique. The probe can spot "hot" points very easily, but points don't necessarily radiate significantly.

Let me give you an example. I was looking for a particular emission source with a near field probe. The source was quite clearly the calibrator pins. The probe showed that they were much more active than any other part of the instrument. And yet when they were removed, complete with their wiring assembly, the emission was exactly the same.

A simple theory for this result is as follows: An RF antenna consists of a current flowing in a length of conductor. The near-field probe detects the current, but not the length. A large current in a very short conductor can be detected by the near-field probe as a major emission source. However, a small current in a long conductor can give much greater emission. If the conductor is much shorter than the (free space) half-wavelength of the stimulating frequency, the radiated power, for a given current, increases as the square of the length of the conductor. In antenna terminology, the ***radiation resistance*** of the wire increases as the square of its length. Furthermore, if the current is only due to the capacitance of the wire, the capacitance can increase almost proportionately to the wire length, making the radiated power increase as the fourth power of the wire length. Given that radiated power in the far field is the product of the **E** and **H** fields, both of these fields can individually increase as the square of the conductor (antenna) length.

In the early stages of trying to locate emission sources you may go around trying various things and getting nowhere fast. It is worth trying the "random" technique because it is the fastest way of finding simple problems. If this method doesn't work then you need to try more drastic and systematic solutions.

Here is a typical problem. You are working on a completely new equipment case design. There is emission at 500 MHz which doesn't seem to be localised. You try copper tape here and there and the difference is never more than a few decibels. You are looking for more like 20 dB improvement. Some copper tape positions even make the emissions worse! Very inexplicable. One possible answer is that there are multiple sources of this emission and you are getting cancellation at some points. Thus reducing one source can make the total emission in a certain direction greater. This is the same problem that you get with any noise debug work.

One solution is to keep adding the modifications that should make improvements and leave them in place. For example there is a seam with spot welds down it. You tape over the seam and the difference is not more than 1 dB; this is almost certainly within the measurement repeatability. Leave the tape on until you have achieved the desired shielding performance. Perhaps the signal level is now down 15 dB below its previous level. *Now* remove the tape. If this tape is having any effect, now is the time it will reveal itself. This makes for positive results and you can be very confident that you are spending money wisely, adding more spot welds, or screws, or whatever, only where necessary.

Another trick is to do modifications many at a time. If you do several sensible modifications and none work you will have saved a lot of time. By 'sensible modifications' I mean ones like taping over holes and slots, removing I/O cables &c; things that should produce an improvement. A less sensible modification would include things like removing a screw or deliberately isolating a piece of metal work. These may make the emission smaller on this one test unit, but there is probably some cancellation effect going on and the result will not be consistent. [I can't say for sure that this is an incorrect modification because you could be isolating a signal source from its radiating antenna. Certainly this sort of modification needs more attention given to it and should not be done along with a whole batch of others.]

These standard approaches still may not give a solution. Now you are getting desperate. The product has to be shipped {sold} but it is still not working. The boss wants to know when it will be done. The solution may seem a bit drastic but it has to be done. You need to know what is causing the problem and you need to be *certain* about it. My technique is to *eliminate* possible causes.

Let me give you an example from a scope design. This particular product had an LCD display outside the enclosure and a keyboard outside the enclosure, although the inside of the keyboard area had been metallised to reduce the radiation from the keyboard. The emission at a few frequencies [up in the 400 MHz+ area] was too high. I wasn't sure if it was the keyboard or the display and I had fitted ferrites to the cables with no effect.

The solution was to pull off the display and put it *inside* the case. This was a tight fit and involved bits of insulator and lots of tape to hold the display away from all the internal electronics. No effect. I put the keyboard inside the instrument as well. Then I taped all the gaskets down to the case. [I had a case with zinc + clear passivation finish because I knew this was not going to be an easy job.] The instrument was all taped up and looked a real mess and it was still over the emission limit!

That meant it had to be the power supply. That was the last thing left. I had the power supply apart and got a pair of feedthrus for the mains wires. I could not find any that were X-rated [see X2 capacitors in the capacitor chapter] so I fitted the best I could find and hoped that they didn't blow up before the test was completed. No change! There was apparently nothing left. The box was sealed up solid metal. It turned out to be the handle that was acting as an antenna. On closer inspection the handle was connected to the internal chassis rather than the outer shell of the case. Hence the metal handle was poking through the RF enclosure. This was not at all obvious because the internal chassis was connected to the outer cover by gasket that was not more than 2 cm from the handle mount. This small length was enough to ruin the shielding. Routing the gasket around the handle mounting boss solved that problem.

A similar problem occurred with a floppy drive which although flush {in line} with the plastic front panel, was actually protruding 2 cm through the RF enclosure. Again this had to be gasketed to the chassis to prevent a particular (>500 MHz) emission spike.

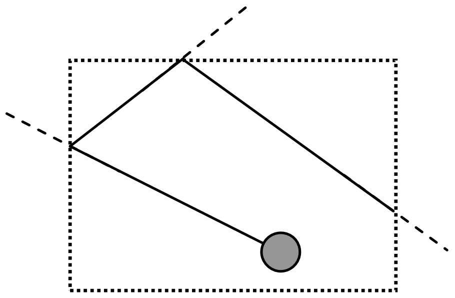
Having had extensive experience with EMI fixing, I was able to get my views implemented on the next range of scopes. The case design was considerably improved. The front panel was difficult. There was a metal plate under the plastic trim, but the plate was not intimately connected to the main chassis. The plate was holding the switches and was useful as a sink for the spark discharge immunity tests. [In these tests a spark discharge gun up at ±15 kV is brought up to the front panel and let loose. Having a definite path to ground for these transient currents is essential.]

I made sure that the plate was grounded to the chassis via a screw. That was the best I could achieve. I would have liked to ground the plate to the chassis at multiple points around the edge, but I could see no simple way of doing so. But then again there was no radiation coupling to this plate so it should have been alright The front panel moulding was done and the finished instrument failed its emissions. This 30 cm long plate was somehow getting an RF signal on it and it was a very effective UHF radiator.

This was a nasty problem to fix. The metal plate was on the outside of the plastic front panel trim so there was no easy way to bond it to the chassis. The final solution required modifying the mould tool so that the panel could be screwed through the plastic moulding and made to clamp pieces of copper tape. The copper tape then made contact with some EMI gasket. A very unpleasant solution brought about by the overall strategy of the case design. It is all very well not 'liking' a particular design of case, but you do need the confidence to be able to say that a particular approach is not workable. The trouble is that you never get to see when you are right! If a particular part of the design works correctly, it is not highlighted. It is the bits that cause trouble that get all the attention.

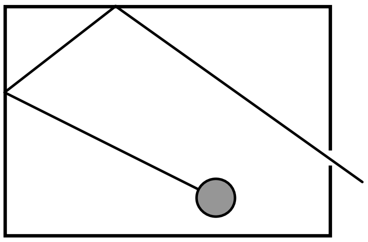
One last point on this experimental stuff. It is quite possible to seal the box up better and yet get a stronger signal registering. One reason is cancellation, and this has been mentioned before. The other reason is more subtle. There are essentially two ways of measuring the power radiated from a piece of equipment. One method is to place the equipment in a fully screened metal chamber and measure the power which is fed into a receiving antenna. The idea is that all power radiated by the equipment will eventually be absorbed by the antenna because it has nowhere else to go. This is the total radiated power of the equipment.

FIGURE 20.3D:



EMC standards measure peak directional radiated power. The equipment is measured from all angles to find the strongest emission; this is the recorded test value. Consider a laser inside a shiny box with holes in it. The laser beam bounces around the box and escapes from many places. The intensity of the beam in any particular direction is therefore not very great.

FIGURE 20.3E:



With the box well sealed up, the laser beam bounces around inside the box and comes out unattenuated in a strong directional beam. This box is therefore *worse* on the peak directed radiation test, despite being "better" screened. In fact the shielding effectiveness of an enclosure can become negative due to resonance.

This resonance effect has been demonstrated both theoretically and experimentally, with shielding effectiveness values as bad as –10 dB to –20 dB.[2] In other words, the radiated field can be as much as 10× *larger* when the case is fitted! In a real situation the resonance is likely to be damped by the innards of the equipment so that the worst SE figures are not usually negative. However, after spending lots of money on a metal case, nobody is going to be happy with a shielding effectiveness below 10 dB. It can be very disappointing to do lots of good work shielding the enclosure only to find that the net result is actually worse than before. It just means you have even more work to do. Don't be disheartened!

In order to minimise the amount of remedial {corrective} work after an enclosure design has been made and tested, you need design guidelines. Specifically you ought to be able to estimate the shielding effectiveness of an enclosure design before you make it. Unfortunately the shielding effectiveness of any particular aperture is necessarily dependant on the size of the enclosure. Hence any formula for shielding effectiveness which does not take into account the size of the enclosure is necessarily incomplete.

Here is a standard text book formula for the shielding effectiveness, *SE*, of a series of holes which are close to each other:

$$SE\,(\text{dB}) = 20 \cdot \log_{10}\left(\frac{\lambda}{2d}\right) - 10 \cdot \log_{10}(n)$$

λ is the wavelength of radiation being considered
d is the diameter of the hole
n is the number of holes

This is based on plane wave shielding theory. It looks very neat and 'plausible'. It is stated as being valid only for $\lambda/2 > d$, in other words for small holes. The term $10 \cdot \log_{10}(n)$ is very easy to justify in power terms. If there are n holes, then the emitted power will be n-times greater. The formula seems self-consistent and text books love to give graphs of the formula, apparently adding weight and respectability to their argument.

With a bit of re-arranging the formula becomes $SE\,(\text{dB}) = 20 \cdot \log_{10}\left(\frac{\lambda/2}{d\sqrt{n}}\right)$

The holes give an open area of $A = n \times \pi r^2 = \frac{\pi}{4} \times nd^2$

[2] M.P. Robinson and others, 'Analytical Formulation for the Shielding Effectiveness of Enclosures with Apertures', in *IEEE Transactions on Electromagnetic Compatibility*, 40, no. 3 (Aug 1998), pp. 240-247.

Substitute into the shielding effectiveness formula $SE\,(\text{dB}) = 20 \cdot \log_{10}\left(\dfrac{\lambda/2}{\sqrt{\dfrac{4A}{\pi}}}\right)$

According to this formula, there is no improvement to be had by making the holes smaller for the same open area. This is demonstrably incorrect (experimentally).

There is a correction to the above formula that is also stated in the same texts. This correction is for the ***waveguide beyond cutoff*** effect. I first encountered this practically when I was trying to improve the radiated immunity of a scope. I had constructed a spark discharge device from an automotive ignition coil, a power transistor and an oscillator. This contraption generated a healthy continuous series of sparks to interfere with the victim scope. Unfortunately I immediately got complaints from my neighbours in the lab. Their computer monitors were going haywire as a result of the spark discharges!

No problem. I got the lab technician to put the spark gap at the bottom of the empty catering-size coffee tin which had just become available. This was a conducting cylinder 21 cm in diameter and 23 cm tall, closed at one end only.

I tried my ‘directed energy weapon’ on the victim scope and it had no effect. The disturbance would not propagate down the length of the coffee tin. The spark was still sparking in a healthy manner, but no significant interference came out. Defeated, I retired gracelessly to my text books for an answer. I found it in a microwave text book. I wasn’t working at microwave frequencies, so how was I supposed to know to search a microwave text book? The answer is that a 21 cm diameter circular waveguide has a ***cutoff frequency*** of around 800 MHz. Below this the waves are severely attenuated. This effect is included in EMC texts as the ‘waveguide beyond cutoff’ formula:

$$\boxed{SE = 30 \cdot \frac{t}{d}\ \text{dB}}$$

t is the material thickness and d is the maximum hole opening.

This formula is only appropriate below the cutoff frequency of the waveguide. The cutoff × size product is 17.6 GHz·cm for circular waveguides.

1 cm	→	17.6 GHz
2 cm	→	8.8 GHz
10 cm	→	1.76 GHz

The waveguide-beyond-cutoff correction term is not sufficient to account for the problem with the hole size in the shielding effectiveness formula, however. Nevertheless, it is an important design guide for shielding.

If an entrance to a screened enclosure is given sufficient depth, then a negligible amount of signal will pass through the opening. For example, if you make a conducting tunnel into a screened room with a diameter of 0.5 cm and a length of 3 cm the formula predicts a shielding effectiveness of 180 dB. Another factor will therefore be the dominant shielding limit and this tunnel will have a negligible effect. This tunnel shielding will be valid up to over 30 GHz.

Such a tunnel has many uses for screened enclosures:

- Allows the passage of fibre-optic control cables.
- Allows switches to be activated by non-conducting push-rods outside the enclosure.
- Allows non-conducting rotatable shafts to be inserted into the enclosure as a variable resonance feature.

The only thing I would say is that *you must not put a conductor through the waveguide beyond cutoff* or you will have made a coaxial conductor; this will be an excellent path for RF energy to both enter and exit the screened region.

The waveguide beyond cutoff phenomenon means that greater attenuation is achieved when an aperture has depth. Thus a fan guard made from an extruded honeycomb is considerably better from an EMC point of view than a flat [2D] mesh.

All in all, the existing formulae for real-world shielding effectiveness are at best ineffective, and at worst unhelpful and misleading. There is great practical benefit in making apertures smaller. I have therefore developed my own empirical formula as a starting point for a design. This is based on work done at 418 MHz on the test box described previously, using many different hole and slot patterns. The experimental technique did not include checking for resonances in the box and therefore could hardly be considered as rigorous. However, the formula is certainly better than the plane wave formulae and it is what I use as a *starting point* in my designs.

$$\boxed{SE = 207 - 40\cdot\log_{10}(d\cdot f) + 5\cdot\log_{10}\left(\frac{d}{h}\right) - 12.5\cdot\log_{10}(n) + 30\times\frac{t}{d} \quad \text{dB}}$$

f is the operating frequency in MHz

d is the maximum linear opening of the aperture in mm.

t is the material thickness in mm.

h is the height of a wide aperture in mm; $h \le d$

n is the number of apertures.

d is a bit tricky. If you have a rectangular aperture then *d* is not the length of the longest side; it is the length of the diagonal. *h* is also tricky. It represents the diameter of a circular aperture, and the height of a wide slot. *h* gives a greater shielding effectiveness figure when a rectangular aperture is shrunk down to a slot.

Suppose the aperture is a long slot. If you electrically connect a fine piece of wire over the middle of the slot you now have two slots of half the width. From a power point of view you would not expect any significant change. The “standard” formula suggests a 4 dB improvement. My formula predicts a 7 dB improvement and I have measured a 10 dB improvement.

You should apply this formula to all apertures on a piece of equipment on a face-by-face basis. In other words, look at the front face and calculate its shielding effectiveness, then look at the left side &c. If there is more than one type of aperture on the face then you have to sum the shielding effectiveness values in a particular way:

a) Make the shielding effectiveness values negative.
b) Convert the –SE values to power ratios.
c) Sum the power ratios.
d) Convert the result back to decibels.
e) The result when made negative is the resultant shielding effectiveness.

As a check, make sure that the resultant SE is less than or equal to the lowest of the shielding effectiveness values being combined.

***EX 20.3.1:** An instrument has a back panel with two screws holding the cover in place. Effectively there are 4 parallel slots each 9 cm long. The cover fits well, but the joint over these slots cannot be guaranteed. You estimate that the slot opening cannot be larger than 0.3 mm. It seems to be about 1 cm deep because of the folds. There is also a hole pattern of 600 off 6 mm diameter holes on this same face. The panel is 1 mm thick. You are looking at 400 MHz signals.

a) What is the shielding effectiveness of the slots?
b) What is the shielding effectiveness of the holes?
c) What is the combined shielding effectiveness?

You may not have noticed that I am summing groups of apertures as radiated power, but individual apertures within a group at a slightly higher rate than that. Thus if you were to calculate on the basis of individual holes and then sum them at the end, you would get a slightly larger SE figure. Remember that this formula is only an approximation. Apply the formula to all of one type of aperture on the equipment face in one go. You could separate out horizontal slots and vertical slots on a face, in order to account for polarisation of the waves, but I would not bother with that.

It is difficult to say how much shielding effectiveness you should aim for. As a starting point I would say that you should aim for at least 20 dB – 30 dB. I would also say that there is not much point in trying to get >100 dB; you are extremely unlikely to be able to achieve such a performance with a single barrier.

20.4 Conducted Emissions

For frequencies up to 30 MHz the test method for radiated emissions is actually by measurement of the RF signal on the mains lead (or indeed any other long connection leads). It is recognised that at frequencies up to 30 MHz the dominant radiation mechanism is by the use of the mains lead and the mains distribution system as an antenna. Frequencies up to 30 MHz are therefore known as 'conducted emissions', but the reality is that this is a measurement method, not an interference mechanism. The mains immunity tests are done at hundreds or thousands of volts, whereas the conducted emissions limits are in the region of 56 dBμV (=631 μV).[†]

The main culprit {guilty party; source} for conducted emissions is switched-mode power supplies. Unless well filtered, they will drive signal back into the mains circuitry and cause an unacceptable amount of emission. Switched-mode supplies usually involve inductive components [inductors and transformers]. If these are of poor quality, or poor design, they will spread magnetic flux around the inside of the equipment. Such an intense field is very hard to shield because its ***wave impedance*** is so low.

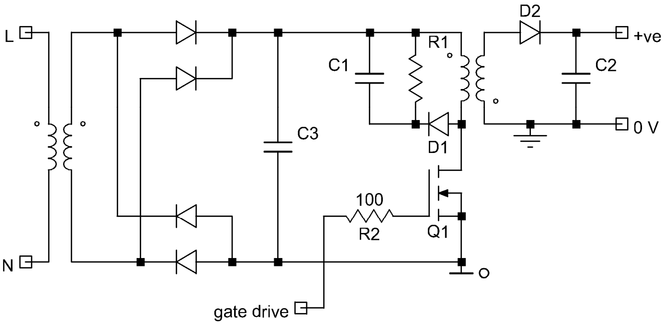
The first thing to do is to minimise the sources of the radiation. This is done by selection of circuit topology, selection of components and careful PCB layout. For

[†] EN55022:1998 Class B at 150kHz

example, if you use axial chokes {inductors} then the open magnetic circuit will just 'spray' magnetic flux everywhere. Toroidal chokes are much better at keeping the flux to themselves. If you are using ferrite transformers then put heavy copper bands around the outside of them. This again reduces the leakage flux, but only if the band is sufficiently conducting. Use a heavy copper band and overlap the joint substantially before soldering it together. It is very easy to 'prove' that such a band is ineffective if you use a thin copper tape and do not make a good joint.

Here is a simplified flyback converter powered from a mains transformer. This is a simple power supply with a wide range of input voltage. It is a flyback converter because of the phasing of the switched-mode choke. Look at the dots marked on the windings. If the dot on the output winding were reversed, this would be a *forward converter*.

FIGURE 20.4A:



Q1 is the main switching device. D1, C1 and R1 are a ***snubber*** circuit to prevent Q1 being subjected to over-voltage spikes. Leakage reactance in the transformer causes excess power dissipation in the snubber.

***EX 20.4.1:** Draw the high speed current paths.

a) When Q1 switches on.
b) When Q1 switched off.

Hint: These are *Maxwell circulating current* paths; that is, they start and end at the same point.

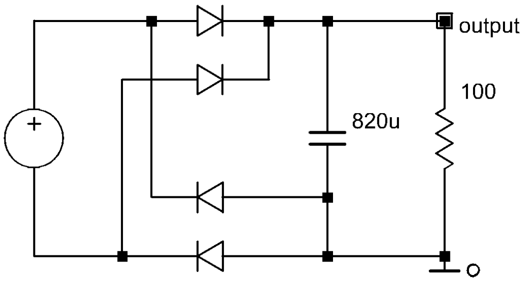
Unless you can draw out the high speed current paths you will not be able to lay out the PCB to minimise magnetic field emissions from these parts. In this configuration it is the switch-off surges that will be the worst. If you take the case of the converter running in discontinuous-mode [the flux in the transformer is allowed to drop to zero on every cycle] the current at switch-on is (ideally) zero. With ideal components the current in the transformer will ramp up linearly with time. When the FET is turned off this current initially tries to continue through the snubber, and in a transformed state through the secondary circuit. Thus the current through most of the initial current path is shut off from the peak value to zero in a short space of time. This is a fast current change and therefore a high frequency emission source.

20.5 Harmonic Currents

There is a requirement to minimise the harmonic currents drawn from the national power grid. If a piece of equipment has a low ***power factor*** then it is taking more current from the mains than it should for its power consumption. Historically poor power factor has been due to large inductive motor loads. These inductive loads have been compensated for in industry by the use of *power factor correction capacitors* (large banks of capacitors) or by *synchronous compensators* (synchronous motors with no mechanical load). The reason for this compensation by industrial users being that factories can be charged for having a poor power factor by the electricity supplier.

The electricity supplier wants to charge for the poor power factor because the extra current is producing extra I^2R {copper} losses in the distribution network. Somebody has to pay for this extra power loss. There is another problem associated with power factor, however. If there is a lot of electronic equipment connected to the mains network, although each individual current waveform is not critical, the mean of all of them is important.

FIGURE 20.5A:



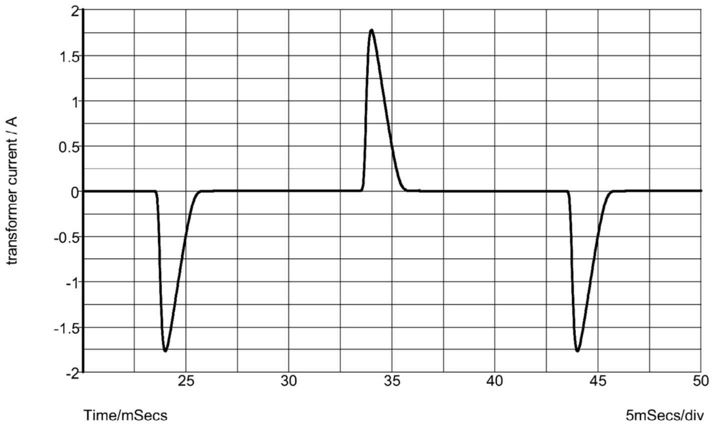
Simple power supply designs consist of a mains transformer, a rectifier, a large reservoir capacitor and a load.

This is a simple simulation model of a full-wave (bridge) rectifier running from a transformer (floating voltage source).

There is voltage ripple at the output due to the finite size of the reservoir capacitor and the load. Any circuitry, such as a voltage regulator, has to work down to the bottom of the ripple voltage. Ripple on the output waveform is highly undesirable. However, look at the current drawn from the transformer.

FIGURE 20.5B:

The current waveform is very 'spiky'. For a given mean current drawn by the load, the RMS current through the transformer windings is increased by using a larger reservoir capacitor. The diodes can only conduct when the voltage at the transformer is larger than that at the load. Hence the requirement of low ripple voltage at the output gives rise to a high ripple current in the transformer and the diodes.



This problem is one of cost and size as far as the designer of the equipment is concerned. There is a choice to be made about how much ripple voltage is acceptable at the output. However, when you look at the situation from the point of view of the electricity supplier, you see an entirely different problem. There was an increasing amount of this electronic equipment being connected to the mains. Since it all had this type of circuitry, there was an ever increasing load towards the peak of the 'sinusoidal' mains supply. This had the effect of flattening the top off of the mains waveform and generating excess power dissipation in the power distribution network.

The solution has been to require that equipment draws a limited amount of harmonic current. The theoretical argument is that harmonic current drawn from a sinusoidal supply is not supplying any power and therefore only produces losses. Actually this argument is flawed because the mains waveform does get distorted and it is therefore possible to supply power at the harmonics. In fact as far as equipment manufacturers are concerned, the ideal mains waveform would be trapezoidal, not sinusoidal. This would reduce the size of smoothing capacitors required, reduce the peak current in the rectifier diodes, and improve the efficiency of the power supplies. As far as the electricity supplier is concerned, however, the voltages are being stepped-up and down through transformers, so a trapezoidal waveform would produce more iron losses {core losses}.

As the situation stands now, there is legislation {law} that requires us to keep the harmonic currents below certain limits. As a concession to very inexpensive low power equipment, there are no limits for ≤ 75 W loads (EN61000-3-2:2000).

The question therefore arises as to how one should design the power supply to meet these new requirements. There are basically two solutions: a big series inductor, or a power factor correcting 'front-end' on the power supply. The big series inductor is not a nice solution. It is a mains frequency inductor and is therefore large and inefficient. On the other hand, the power factor corrected front-end can be even more expensive than the simple inductor solution. Just how expensive and how inefficient these things are is all a matter of technology and design.

The power factor corrected front-end is basically a switched-mode supply that draws current from the supply in proportion to the instantaneous supply voltage. It therefore looks like a resistive load if you average out the high frequency switching currents. If the front end switched-mode supply directly fed the output capacitors then it would not be possible to achieve low amounts of ripple unless huge output reservoir capacitors were used. Hence power factor corrected switchers tend to create a relatively high-ripple intermediate DC voltage, which is then fed to a second stage of switched-mode regulation.

Given that electrolytics cost more according to their physical size, and that you can get more energy stored per unit volume in high voltage electrolytics, it is usual to run the intermediate stage at several hundred volts, rather than a few tens of volts.

EX 20.5.1: An off-line {directly connected to the 50 Hz mains} switched-mode supply running at 30 kHz has been designed to look resistive at its front end. It is supplying power directly to a single output capacitor at 20 V and there is a 4 A load on the 20 V rail. Estimate the required capacitance of the output capacitor in order to keep the output ripple voltage below 100 mV ptp. Neglect the power loss and ESR or ESL in the output capacitor. State any assumptions made, but don't worry about tolerances, selecting E6 capacitor values or any other practical considerations.

A two-stage switcher, which uses a power factor front end feeding an intermediate DC voltage, can use much smaller capacitor values for the same amount of output ripple. Note that the ripple current rating of the output capacitors will be unchanged, except for the fact that ripple current ratings at 10 kHz can be anything from 10% to 100% larger than the 100 Hz ratings of the same capacitor.

When measuring harmonic currents you might assume that if you used a harmonically distorted supply and the harmonic currents were within the limits specified in the standard, then the equipment would be ok. That is **wrong**. If the load is resistive then the previous statement is correct, but the whole point of measuring the harmonic currents is because the load is not resistive.

An ordinary industrial mains supply will look like a flattened sinusoid. The tops of the waveform are smoothed off because lots of equipment is conducting near the peak of the sine wave. This is the very problem you are trying to avoid on the newer equipment. If you look back at the rectifier and capacitor circuit earlier in this chapter you will understand that a flat topped waveform will conduct into the capacitor for a greater part of the cycle. There will be less ripple voltage on the capacitor and the current in the diodes will cover a larger portion of the mains cycle. (Older books would say that the *conduction angle* of the diodes is larger.) Less current surges means less harmonic content.

I have measured the difference a pure supply made to the current waveform using a scope as the load. I was surprised to find out that the repetitive peak current drawn by a power supply doubled when it was fed from a harmonically pure mains source!

20.6 ESD

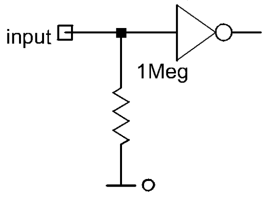
Electro-**S**tatic **D**ischarge presents two problems for the designer: one is damage and the other is circuit mis-operation.

Quite clearly the first thing you have to do is to prevent damage. After that you can then make sure that the circuit operates as intended as well. The EMC standards give levels of severity for the discharge from a simulated human body model. The human body model is charged up to perhaps ±8 kV and discharged into the circuit under test. When you get up to these levels, ordinary components are not rated for the application at all. Fortunately, on common digital interface connections, the major manufacturers have come up with I/O chips that are now rated up to the full ±15 kV (air-discharge) limits. Non-standard I/O ports still require protection networks to be designed though. How do you protect your polarised bean counter[†] input from static discharges? The whole topic of input and output protection networks needs careful consideration.

I like to think of equipment users as a sub-species of *homo sapiens*; let's call them *homo stupidus*. If there is a sensitive input on your equipment then *homo stupidus* will wire it up to the mains. If there is a sequence of actions that needs to be followed to protect the equipment then *homo stupidus* will do the sequence backwards, or sideways, or any way but the correct way. In essence you have to protect your equipment from *homo stupidus* because if the equipment breaks, *homo stupidus* will blame the equipment and will want it fixed for free under warrantee.

You will notice that some expensive DMMs protect their ohms ranges against 240 V AC inputs. You will also notice that most general purpose scopes protect even their most sensitive ranges up to 400 V peak as well. *Homo Stupidus* has been with us for a long time, and now that civilised society has minimised natural selection, his numbers are not being diminished!

FIGURE 20.6A:



In this respect, ESD is no more of a problem than any other spurious input. Consider a high impedance logic input, a CMOS gate for example. This is a logic interface to the outside world 'designed' by a digital engineer. At least there is a defined logic level when the cable is unplugged. Some 'designers' might not even include the resistor!

[†] Invented product type

High impedance circuitry such as CMOS needs to be 'told' what DC level to go to, all of the time. If it is left to float then it will do unpleasant things like cause the gate to draw excessive power supply current, to produce spurious logic signals, to create interference to other parts of the circuit, and generally to make a nuisance of itself. The first thing you have to decide is what you need to protect the circuit against. What are the likely things that *homo stupidus* could connect to the input?

This is a difficult question to answer, but it is a vital first part in the design. If you are dealing with an output then you would assume that it could get inadvertently short-circuited for an indefinite time. If there were any power rail connections in the same connector then maybe these could get shorted to the output as well. The same thing applies to an input. The very minimum amount of protection would be to assume that any of the other pins in the connector could get wired to this input.

On a certain type of temperature controller there used to be a big terminal strip on the back where the I/O connections were made. They were all the same type of terminals [which in retrospect seems pretty silly] so it was assumed that somebody could accidentally wire the mains [230 V] up to the thermocouple inputs [which expect only a few millivolts]. Protection against mains is quite a common requirement on an I/O pin.

Logic gates are not usually blown up by "voltage" so much as current.. It is usual for there to be diodes from the input pins to the power rails. These may be parasitic diodes due to the process, or they may be deliberately added to provide a better termination for fast signals. In any case it is not fatal to forward bias these diodes. However, it is fatal to put too much current through these diodes. Depending on the technology, a forward bias current of 10 mA in the protection diodes may be acceptable. Your task is therefore to limit the current to less than this [or to some other amount specified by the manufacturer.]

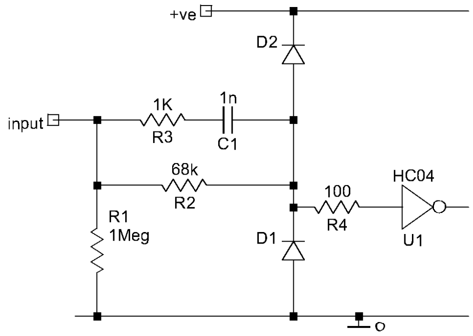
No problem; a 240 V sine input peaks at 339 V. Just neglect the transients that occur on the mains because this overload condition is the transient and you are not expecting it to be applied for very long. The logic gate is on a 5 V rail say, so you need a resistor of (339−5)/10m = 33K. As far as the resistor is concerned it is virtually connected straight across the mains, so the dissipation in it is ≈1.7 W. It is therefore quite common to see large high-power resistors used in the protection circuitry of inputs.

Suppose you can only fit in a 1 W resistor due to the size; you might then go for a 68K resistor. [I am not too concerned with tolerances in this discussion, although you would obviously take these into account in the actual design.] The circuit is now protected, but it won't work at any speed. After all, if the gate input and connector/track capacitance is say 10 pF then the resulting bandwidth is only 234 kHz. This may be plenty; on the other hand it may have ruined the performance of your system. It is all a question of the system requirements. To speed it up you can put a capacitor across the resistor. If this capacitor is 100× bigger than the total input capacitance, then the signal will not be dramatically attenuated, but the 1% loss has to be considered. You might use a 400 V 1 nF capacitor.

The trouble is that you now have a high speed path straight into the input with a very low source impedance. The input will get destroyed immediately the signal is connected because the inrush current via the capacitor will almost certainly blow the gate. The trick is to put a small resistor in series with the capacitor to limit the transient current, typically something around 1K would be suitable. The bandwidth has now increased by 68× and this may well be acceptable.

Be aware, however, that the gate may not be sufficiently well protected against fast spikes like ESD pulses, which range up to ±15 kV.

FIGURE 20.6B:



This is a typical protection scheme. R3 might be as low as 100 Ω, but omitting it completely is not a good idea. The diodes should be low-leakage small-signal types (eg BAV199) in order to minimise capacitive loading. R1 has been shown at the input, where it sees the full input transient. Putting R1 across D1 is wrong because the input gets attenuated and the input impedance varies with frequency.

R4 is a more subtle component position. The power rails of the gate have not been explicitly drawn, but the diodes are typically connected to these power rails. Without R4 the input overload current is shared between the internal and external clamp diodes. The external diodes are bigger than the internal diodes and therefore their share of the transient current is larger. However, the shared ratio is not well defined and the gate can still get blown up. Some people use schottky clamp diodes to minimise the current into the gate protection diodes. The problem with this is that schottky diodes, at high surge currents, can be worse than ordinary silicon diodes in terms of their volt drop. R4 prevents excessive gate current.

As an alternative to using a high power R2, this position can be replaced by a PTC thermistor. The idea is that the thermistor is a few hundred ohms when it is cold, but when the input has hundreds of volts applied, the thermistor heats up and the resistance shoots up to >10 kΩ. This approach is not necessary for digital inputs, but may be necessary for analog inputs. The bias current related errors when using large protection impedances can ruin the measurement accuracy.

It has to be said that using PTC thermistors is very much an experimental design procedure in the sense that it is difficult to calculate the overload conditions in the diodes. You will therefore need to do some testing on the 'finished' scheme to ensure that the diodes are not being over-stressed by the initial in-rush current while the thermistor heats up and limits the current.

Static discharges result from people touching equipment. It is very common for people to get minor spark discharges from their fingers when touching metal filing cabinets and other large metal surfaces. Such charges are generated by dissimilar materials rubbing and displacing charge. In fact just the action of making and breaking contact between two dissimilar materials can give rise to charge separation. Synthetic materials are known to be considerably worse for creating static than natural materials like cotton.

Rather than testing the equipment by dragging your feet on a carpet, the standard method is to use an ESD simulator, also known as a static discharge gun. For the human body model, consider a capacitor of ≈100 pF with a series resistor of ≈100 Ω. The capacitor is charged up to between ±2 kV and ±15 kV, depending on the severity requirement of the test.

Traditionally the test method has been 'air-discharge'. The tip of the ESD gun is charged up to the test voltage and then the whole gun is moved towards the victim circuit until a spark occurs. This is not a very controlled method because the rate of approach of the gun to the victim circuit is important. The more controlled method is a 'contact-discharge'. The tip of the ESD simulator is held in place against the victim circuit (in contact with it) and the trigger is pulled. This gives a more repeatable test for exposed contacts and inputs.

Note that the maximum air discharge limit is currently ±15 kV, whereas the maximum contact discharge is ±8 kV. The standards committees have determined that the destructive energy in a 15 kV air discharge is approximately equivalent to a contact discharge at 8 kV. Thus the designer does not have to invent protection requirements from 'first principles'. The standards committees do this part of the spec and it is only necessary to pick a severity level. Again there are various additional standards which recommend severity levels. The relevant standards give exact test conditions to follow and effectively teach you how to do the tests. IEC61000-4-2 is a widely used standard.

An instrument front panel is a typical problem area. The controls may have plastic knobs, paddles and buttons, but there is internal electrical/electronic circuitry nearby. You have to feel around with the ESD gun (in air-discharge mode) looking for a weak spot. At 15 kV it is easy for the spark to sneak around corners and discharge to a PCB track.

This level of spark will easily wipe-out {destroy} many logic devices; protection is necessary. The simple solution is a pair of diodes to the power rail and a series resistor into the gate I/O pin.

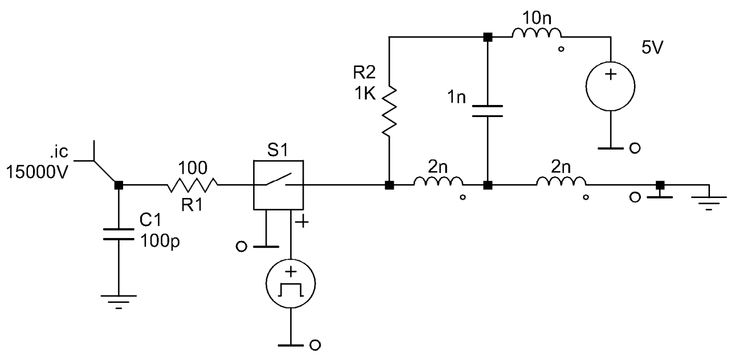
Often shafts of switches and metal front panels are available as discharge paths and you must control the path of the current. You have to develop the ability to 'see' where the current is going to go. I am talking about 15,000 V wavefronts with < 500 ps risetimes travelling around the circuit. If that doesn't scare you in terms of protecting poor little 5 V CMOS logic then you haven't understood the problem!

The charge from the ESD pulse 'wants' to get to the outside faces of the conducting enclosure. This is just your elementary physics classes on electrostatics being brought into play. You would do well to pull out the simplest text you have on electrostatics and see the experimental work that showed that charges reside on the *outside* of conducting surfaces and not the inside. (Faraday's ice-pail; 1843)

The charge needs to spread out over the entire outer surface of the enclosure and it starts out at the point it is injected. You have to look at the injection point and put yourself in the place of the electrons; how would you get to the outer shell of the enclosure in the minimal possible time?

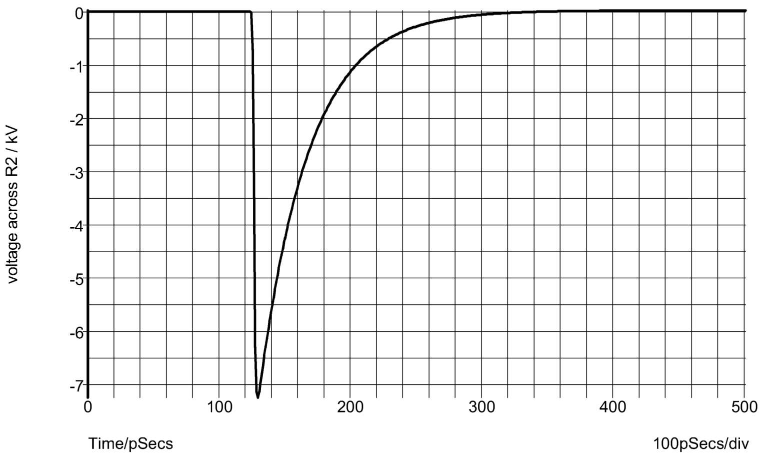
One thing is for sure, if this current flows through a PCB ground plane on which other circuitry is connected there is going to be a problem. I don't care if it is solid ground plane and it is multiply bonded to the case, the voltage gradient across the board will be huge. Just imagine 15,000 V across a 15 cm path. If evenly distributed that's 1000 V/cm. If the discharge is going into the ground plane, and the positive rail decoupling capacitor ground connection is more than a few millimetres away from the ground connection of a logic device, the device will get a positive or negative power surge which may reset flip-flops and therefore cause an unintended change in logic state.

FIGURE 20.6C:



The switch S1 applies the static discharge from the simple human body model C1–R1 into the circuit ground plane. R2 represents the power connections of a logic gate. Regardless of the exact values used, the results are horrifying.

FIGURE 20.6D:



An initial 7 kV reverse voltage occurs across R2 in the short term. The waveform then appears to return to zero on this kilovolt scale. However, closer inspection shows that the waveform then rings between –80 V and +80 V at roughly 45 MHz.

This model is not complicated enough, or accurate enough, to simulate what actually happens, but it is good enough to show that there will be a serious problem with static discharge pulses into a ground plane.

Parallel paths will not help in this instance. You need to separate the ESD current path from the circuitry ground path. This means multiple ground planes or separately grounded metal plates. I can't give you a general construction. Just realise that the RF ESD current path must not flow along the same path as the circuit ground. One useful thing to remember is that the very fast edge will want to go to the outside of any metal enclosure, rather than flowing through the middle. You should use this knowledge to provide a direct path from the ESD injection point to the outer case of the equipment.

Also realise that you only fully understand the problem when you can draw a representative simulation model or equivalent circuit. The nature of the problem is then revealed.

As semiconductor junctions get smaller, the ESD event withstand capability reduces. Semiconductors for microwave and mm-wave uses are therefore extremely static sensitive. The sensitivity gets to the point that despite conductive floors, wrist straps and safe handling procedures, these devices can still get destroyed when the relative humidity is too low. For this reason ESD ultra-sensitive components and devices require controlled humidity environments, the lower humidity limit being set between 45% and 50%.

20.7 European Directives

This is just a selection of a few of the important European Directives affecting the electronics industry. There are also plenty of amending directives not listed.

ATEX Directive: 94/9/EC. Equipment and protective systems intended for use in potentially explosive atmospheres.

EMC Directive: 89/336/EEC. Electromagnetic compatibility.

Low Voltage Directive: 73/23/EEC. Electrical safety.

Machinery Directive: 98/37/EC. Safety for machinery.

RoHS Directive: 2002/95/EC. Reduction of hazardous substances in electrical and electronic equipment.

WEEE Directive: 2002/96/EC. Waste electrical and electronic equipment

The text of these directives, and others, is freely available on the internet, although any particular website given may no longer be available at the time of reading.

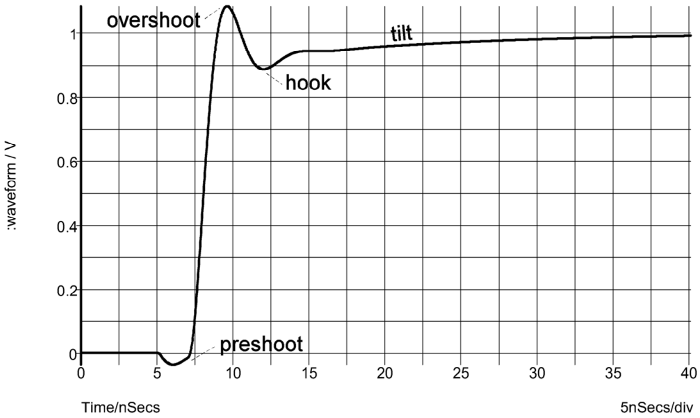
http://www.compliance-club.com/
http://ec.europa.eu/enterprise/newapproach/standardization/harmstds/reflist.html

The Directives are enacted into law in member states by means of country specific *regulations*. For example in the UK, the EMC directive was made law by the introduction of Statutory Instrument 1992 No. 2372, The Electromagnetic Compatibility Regulations 1992.

http://www.opsi.gov.uk/

ENCYCLOPAEDIC GLOSSARY

ABERRATION: If a rectangular input pulse produces any sort of non-rectangular response, then that is *aberration*. if the pulse edge speed is just slowed down then that is not ordinarily considered as aberration.



If instead of heading up to the steady value (tilt) the waveform heads down to the steady value, then this is referred to as *droop*. The *pre-shoot* can be in the opposite direction to the rising edge, as shown above, or it can be in the same direction.

If the pulse overshoots, then dips below the final settled value, the dip after the overshoot is called *hook*. If there is slow upwards tilt, this has been called *dribble-up*.

AFC: Automatic **F**requency **C**ontrol. In a *super-heterodyne* radio receiver, drift of the *local oscillator* may cause the received station to drift out of tune. The AFC circuit is designed to lock the local oscillator relative to the incoming carrier in order to keep the station tuned in.[1] AFC often has its own on/off switch on FM radios. AFC circuitry is the forerunner of the ***phase-locked loop***.

AIR DISCHARGE: Air at normal temperature, pressure and humidity, breaks down if the electric field strength is too high. For parallel plates the requirements for discharge are:

1 mm	4,300 V	4.3 kV/mm
2 mm	7,400 V	3.7 kV/mm
10 mm	30,000 V	3.0 kV/mm

The breakdown is not proportional to distance. Furthermore, sharp points cause the discharge to occur sooner, at more like 1.5 kV/mm.[2]

As an approximation, take the breakdown voltage as being proportional to atmospheric pressure and inversely proportional to the absolute temperature … $V_{BREAKDOWN} > 300\text{V} + \left[P_{bar} \times \frac{290}{T_K} \times 3\text{kV/mm} \right]$

Since irregularities in the electrodes and dust on their surfaces lower the breakdown voltage, keep the calculated field strength below 1.5 kV/mm. For spheres this gives a minimum diameter of 1.3 mm/kV, giving 13 cm minimum diameter at 100 kV.[3] The minimum diameter of a cylindrical conductor can be about three times smaller than that calculated for the sphere at a given voltage. For applications running at tens of kilovolts or more, conductors clearly need to be large and rounded. The alternative is to coat the conductors in a strong insulating material having a breakdown strength well above that of air. This insulation allows smaller interconnections to be made in high voltage applications. It is found experimentally that a positively charged electrode produces corona much more readily than a similar electrode configuration charged negatively with respect to ground.[4] Therefore make the smaller of a pair of electrodes negatively charged to minimise corona.

[1] C. Travis, 'Automatic Frequency Control', in *Proceedings of the Institute of Radio Engineers*, 23, no. 10 (Oct 1935), pp. 1125-1141.

[2] 'Dielectric Strength' in *BR229: Admiralty Handbook of Wireless Telegraphy, Vol 1* (His Majesty's Stationery Office, 1938), p. 175.

[3] E. Kuffel, W.S. Zaengl, and J. Kuffel, '4.2.2 Coaxial Cylindrical and Spherical Fields', in *High Voltage Engineering Fundamentals*, 2nd edn (Butterworth-Heinemann, 2000), pp. 209-214.

[4] E. Kuffel, W.S. Zaengl, and J. Kuffel, '4.1 Electrical Field Distribution and Breakdown Strength of Insulating Materials.', in *High Voltage Engineering Fundamentals*, 2nd edn (Butterworth-Heinemann, 2000), pp. 201-205.

ALLAN VARIANCE [5] … is a measure of the stability of an oscillator, originally developed for the measurement of atomic frequency standards. The absolute mean frequency of an oscillator cannot be known unless it is measured for an indefinitely long time. Thus any measure of the stability of an oscillator is an estimate based on the measurement interval.

The Allan Variance takes successive mean values of the frequency of the oscillator over intervals of duration τ with no 'dead time' between the measurements. The mean of the squared differences of these readings is then used to calculate the variance. Denoting the individual mean readings as $\bar{y}_1, \bar{y}_2, \bar{y}_3 \ldots \bar{y}_N$, the Allan Variance is given by:

$$\boxed{AVAR \equiv \sigma_y^2(2,\tau) \equiv \sigma_y^2(\tau) = \frac{1}{2 \cdot N \cdot \bar{y}^2} \sum_{r=1}^{N} (\bar{y}_{r+1} - \bar{y}_r)^2}$$

Notice that the variance has been normalised by dividing by $\bar{y}^2$ and is therefore dimensionless. The mean value, $\bar{y}$, is measured over some much greater interval than τ.

$$\bar{y} = \frac{1}{m} \sum_{r=1}^{m} \bar{y}_r \; , \qquad m >> N$$

ALIAS: In a sampled data system, an alias is a sampled signal which has a lower frequency than the input signal. An alias will occur when there are less then two samples per cycle of the highest significant frequency contained within the input waveform. On a sinusoidal input signal, the alias will be a sinusoid of a different frequency. For a non-sinusoidal signal, the resulting waveform will appear as a slower version of the input signal, but may be time reversed.

Aliasing can be used to get a higher effective sampling rate on a repetitive signal. It also allows the step response of a sampling system to be measured, even when the sample rate is many times lower than the analog bandwidth.[6]

If $\left(\frac{F_{SIGNAL}}{F_{SAMPLE}} \mod 1 \right) < 0.5$ then the aliased signal will be in the correct time sequence; otherwise the alias will be time-reversed. Time related features on an alias will scale according to the apparent frequency of the aliased signal compared to the actual signal frequency. Suppose a 1.001 MHz square wave is sampled at 1.000 MHz. The resulting alias will be a square wave of 1 kHz. If the risetime on the alias is 9.7 μs then the combination of the system step response and the squarewave edge speed is actually $9.7\mu\text{s} \times \frac{1\text{kHz}}{1\text{MHz}} = 9.7\text{ ns}$. To get a lower effective sample rate just ***decimate*** the data. For example, if the system samples at 10 MS/s, and you want 1 MS/s, just discard the last 9 points out of every block of 10.

ANTENNA FACTOR: An antenna {aerial} placed in an RF electric field produces an output voltage into a defined load, usually either 50 Ω or 75 Ω. The antenna factor is the calibration constant used to convert from electric field strength to output voltage; unless otherwise specified, assume the antenna factor relates to a 50 Ω load. In stating an antenna factor, it is always assumed that the antenna is pointing in the direction of maximum signal reception of an incoming uniform field.

RF electric field strength, **E**, is often expressed in dBμV/m rather than V/m, where it is understood that these are RMS values of sinusoidal quantities. It is convenient to express the antenna factor, ***AF***, in dB/m. $\boxed{E = V_{OUT} + AF}$ when measured in dBμV/m, dBμV and dB/m respectively.

[5] D.W. Allan, 'Statistics of Atomic Frequency Standards', in *Proceedings of the IEEE*, 54, no. 2 (Feb 1966), pp. 221-230.

[6] L.O. Green, 'The Alias Theorems: Practical Undersampling for Expert Engineers', in *EDN* (Cahners), June 21, 2001, pp. 97-105.

(Don't confuse dBm, decibels above a 1 mW power level, and dB/m, decibels per metre.) If the measured voltage at the antenna output is 30 dBµV ($=10^{30/20}\ \mu\text{V}=31.6\ \mu\text{V}$) and the antenna factor is 15 dB/m, the incident field strength is 45 dBµV/m. A larger antenna factor, at a given frequency, means a *less* sensitive antenna.

For resonant dipoles, and wide-band antennas such as log-periodic variants, the antenna factor generally increases with frequency. The reason for this drop in sensitivity is understandable from power considerations. The 'active part' of a wideband antenna at any given frequency has the approximate dimensions of a resonant half-wave dipole. As the frequency increases, the equivalent dipole therefore gets shorter, and less of the incoming radiated power is intercepted by the active part of the antenna. In practice therefore, the antenna factor ordinarily increases with frequency by roughly 6 dB/octave (20 dB/decade) even for well designed broadband antennas. However, at the low frequency limit of operation, the antenna factor also increases by 6 dB/octave of decreasing frequency; in this case the antenna is getting too short to be very effective.

This increase of antenna factor with frequency is seen when the antenna factor is calculated from the antenna's *practical gain* {*realised gain*}, the standard gain multiplied by the Z_0 *mismatch loss* factor, $\left(1-|\Gamma|^2\right)$.

$$\boxed{AF = 20\times\log_{10}\left(f_{MHz}\right)-G_{PdBi}-29.78\quad\text{dB/m}}$$

The frequency is measured in megahertz, and the antenna gain is measured relative to a loss-less isotropic antenna, taking into account the finite reflection coefficient of the real antenna.

Unless otherwise stated, "antenna factor" means the *free-space antenna factor*. Antenna gain, and hence antenna factor, are strongly affected by the presence of nearby conducing surfaces.

Another way of looking at a receiving antenna is to consider its *effective height* (also called *effective length*) in a direction at right angles to the incoming electromagnetic wave.

$\boxed{V_{OUT}=E\times h}$ when measured in V, V/m and m respectively.

The antenna factor is related to the effective height. $\boxed{AF=-20\times\log_{10}(h)\quad\text{dB/m}}$

A dipole needs a ***balun*** to match it to a single-ended 50 Ω system and this balun needs to be included in the antenna factor calibration. A typical antenna factor for a dipole and balun would be –2 dB/m at 30 MHz; scaling by 20 dB/decade gives 18 dB/m at 300 MHz.

For microwave frequencies a horn antenna is used. The antenna factor of a particular wideband horn antenna[†] is given approximately by $33+0.5\times f_{GHz}\quad\text{dB/m}$ over the range 2 GHz to 18 GHz, giving 34 dB/m @ 2 GHz. The VSWR is less than 2 over this range.

In addition to taking into account the tolerances on the antenna factor, the measured signal voltage and the Z_0 mismatch loss factor, the uncertainty in the measured field strength also needs to take account of the mismatch between the antenna and the measuring device (receiver). In decibels this *mismatch uncertainty* is given by:

$$\boxed{\text{mismatch uncertainty} = 20\times\log_{10}\left(1\pm|\Gamma_A|\cdot|\Gamma_R|\right)\quad\text{dB}}$$

where Γ_A and Γ_R are the reflection coefficients of the antenna and receiver respectively. This mismatch uncertainty contribution can be made small enough to neglect by the liberal inclusion of 3 dB pads (attenuators) throughout an RF measurement system.

ANTENNA GAIN … is a slightly misleading term because a passive antenna has no actual power gain. Antenna gain is really a measure of the ***directivity*** and radiation efficiency of the antenna.

The gain of an antenna is always stated relative to a reference antenna. If the reference antenna were an isotropic radiator, the power transmitted by the antenna would be radiated equally in every

[†] ETS·Lindgren model 3117

direction [hence isotropic]. The power flux density (W/m^2) through a small surface of the resulting spherical radiation pattern at a distance *r* would be the transmitted power divided by the surface area of the sphere at that distance. (An ideal sun would be an isotropic radiator.)

$$\boxed{\text{isotropic power flux density} = \frac{\text{(transmitted power)}}{4\pi r^2} \quad \text{W/m}^2}$$

Any real antenna has a directional pattern in space, with a peak radiated signal strength in one particular direction; the ratio of peak power flux density to mean power flux density (both measured at the same distance from the antenna) is called the *directivity, D*.

$$\text{peak power flux density} = D \times \frac{\text{(transmitted power)}}{4\pi r^2} \quad \text{W/m}^2$$

For a lossless antenna that is perfectly matched to the transmitter, the power *delivered* to the antenna is equal to the transmitted power. In general, the transmitter delivers power to the antenna, some of which is accepted and some of which is reflected.

$$\text{delivered power} \times \left(1 - |\Gamma|^2\right) = \text{accepted power} \quad \text{W}$$

$$\text{delivered power} \times |\Gamma|^2 = \text{reflected power} \quad \text{W}$$

The delivered power is often referred to as the *forward power* or the *incident power*. Of the accepted power, some is lost as heat in the antenna structure (inefficiency) and the rest is transmitted, the ratio of these two powers being the radiation efficiency η.

$$\text{transmitted power} = \text{accepted power} \times \eta \ .$$

$$\text{peak power flux density} = \eta \times D \times \frac{\text{(accepted power)}}{4\pi r^2} = G_i \times \frac{\text{(accepted power)}}{4\pi r^2} \quad \text{W/m}^2$$

G$_i$ is the standard power gain relative to an ideal isotropic radiator, denoted by G$_{dBi}$ when expressed in dB. The antenna still has to be losslessly conjugate-matched to the transmitter if the accepted power is going to be equal to the delivered power.

$$\text{peak power flux density} = G_{Pi} \times \frac{\text{(delivered power)}}{4\pi r^2} \quad \text{W/m}^2$$

G$_{Pi}$ is the *practical gain* relative to an ideal isotropic radiator, in other words it is the standard antenna gain multiplied by the *Z$_0$ mismatch loss* factor.

$$G_{Pi} = G_i \times \left(1 - |\Gamma|^2\right) \qquad \text{[when not expressed in dB]}$$

Note that no power is actually dissipated by this reflection, despite the use of the synonymous terms *Z$_0$ mismatch loss* factor, *reflection loss factor*, and *reflection efficiency*.

For optimum power transmission, a lossless reactive *antenna tuning unit* would be needed to *conjugate-match* the antenna to the transmitter, thereby eliminating the $\left(1 - |\Gamma|^2\right)$ 'loss' factor. A receiving antenna has the same loss factor.

The uncertainty in an antenna gain calibration is at best around the ±0.5 dB level. By the time you take into account the uncertainty in the cable loss, *mismatch uncertainty*, and site attenuation variations, it is likely that the uncertainty of an RF field strength measurement will be around ±3 dB, even using good equipment. Contrast this with DC voltage measurements where ±10 ppm measurements are easy to reproduce.

APERTURE (of an antenna): If an antenna is placed in the path of a uniform plane electromagnetic wave, it will absorb some of the incident radiation. It is sometimes convenient to think of all of the incident radiation in a certain cross-section of the wavefront being absorbed, and that therefore the antenna has an effective “capture area” known as the *effective aperture*, A_e. The term *aperture* is derived from microwave horn antennas, where the size of the open horn is its physical aperture.

$$\boxed{A_e = \eta \cdot D \cdot \frac{\lambda^2}{4\pi} = G_i \cdot \frac{\lambda^2}{4\pi} \quad \mathrm{m}^2}$$

The terms in this formula are not expressed in dB.

λ is the free-space wavelength of the electromagnetic radiation in metres

D is the directivity of the antenna (a dimensionless number)

η is the radiation efficiency

G$_i$ is the gain (a dimensionless number) relative to a lossless isotropic antenna

A half-wave dipole antenna can have arbitrarily thin conductors and yet it still has an effective aperture of $0.13\lambda^2$, meaning that the effective captured area could be considered as extending out to a distance of 0.13λ on either side of the conductors, forming a rectangular area $\frac{\lambda}{2} \times 0.26\lambda$. On the other hand, a Yagi-Uda antenna can have the same frontal area as a dipole but with 10× the directivity, and hence 10× the effective aperture.

APERTURE DELAY: On a sampling system [such as a sample & hold or an ADC] aperture delay is the time between the sampling clock reaching the switching threshold and the analog input being sampled. On multi-converter systems the sampling times may need to be synchronised. The aperture delay difference between the converters is known as *time skew*, often abbreviated to *skew*.

In general, aperture delay varies with the instantaneous value of the applied input signal, causing increased harmonic distortion at higher signal frequencies. It is likely that a system with a large aperture delay will also have a large ***aperture jitter***.

APERTURE JITTER … is the change of ***aperture delay*** on a set of sampled data points due to noise. Aperture jitter is one cause of increased signal degradation when the signal frequency is increased. The formula for the signal-to-noise ratio due to jitter on the sampling clock is $SNR_{dB} = -20 \cdot \log_{10}\left(2\pi \cdot f_{signal}\ \delta\ t_{RMS}\right)$. See the appendix for the derivation.

Signal Frequency	Sampling Jitter	Signal to Noise Ratio	Effective Number of Bits
1 MHz	1 ns (RMS)	44 dB	< 7.0 bits
10 MHz	100 ps (RMS)	44 dB	< 7.0 bits
10 MHz	10 ps (RMS)	64 dB	< 10.3 bits
100 MHz	5 ps (RMS)	50 dB	< 8.0 bits
1 GHz	0.32 ps (RMS)	54 dB	< 8.6 bits

Aperture uncertainty is not the same thing as aperture jitter. For any particular sampling device there will be uncertainty as to exactly where the sample will be taken due to propagation delay variations from device to device. For any particular device, the variation with time of the aperture delay will be somewhat less than the aperture uncertainty. The aperture jitter is therefore a smaller region within the aperture uncertainty window.

APERTURE TIME … is the effective time that the sampling gate is open in a sample & hold. As the aperture time is increased, the bandwidth of the sample & hold is reduced, the high-speed signals being averaged out to some degree. If the sampling system is considered as an integrator, it should be clear that when the aperture time is equal to the period of the input signal, the output will be zero.

ARC: When an electrical discharge occurs through air, the resulting arc has a negative AC resistance characteristic. As the current is increased, the AC resistance of the arc decreases. In power system engineering the resistance of an arc (short-circuit fault) is approximated by:

$R_{ARC} = 0.044 \times \dfrac{V_{NF}}{I_F}$, where I_F is the fault current and V_{NF} is the supply voltage with no fault.

This simple approximation is used for power systems up to 110 kV. Above that the multiplier is changed from 0.044 to 0.022. Multiplying the arc resistance by the current in the arc shows that the voltage across the arc is approximately constant, regardless of the fault current. For smaller arcs, of the order of a few millimetres, increasing the current causes a definite reduction in the voltage across the arc.

AVAILABLE POWER … is the maximum power that can be extracted from a source. This is achieved by *conjugate matching* of the load impedance to the source impedance. If the source has an impedance of $R+jX$, the conjugate-matched load is $R-jX$. The reactive parts of the impedances are equal but opposite; the resistive parts are equal.

AVERAGE … is more of a qualitative word than a definite measure; as such it is best avoided and replaced by a more exact term. The most usual average measure is the *mean*, also known as the *arithmetic mean*. Unless otherwise specified, when the term *average* is used, the *mean* is what was intended. The mean value of a discrete variable *x* is the sum of all the individual *x* values divided by the number of *x* values used.

$$\text{mean of x} = \langle \text{x} \rangle = \bar{x} = \frac{1}{N} \cdot \sum_{n=1}^{N} x_n = \frac{1}{N} \cdot \sum_{n=0}^{N-1} x_n$$

All these notational forms represent the same thing. The angle brackets are sometimes used specifically for a time averaged value, the data points being acquired at uniform time intervals. Other types of average include the *median*, the *mode*, the *geometric mean* and the *harmonic mean*.[7]

The *median* is found by sorting the *x* values into ascending order and taking the mid *x* point. If the data set has an even number of elements then the mean of the two middle elements is used. The median is used to give an 'average' which tends to neglect the extreme values.

The *mode* of a dataset is the value that occurs most frequently.

The *geometric mean* of a set of *N* numbers is defined as the *N*th root of their product.

$$\text{geometric mean} = \sqrt[\text{N}]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_N} = \left(\prod_{n=1}^{N} x_n \right)^{1/N}$$

The capital-π notation means that all the terms given are multiplied together.

The *harmonic mean* is the reciprocal of the (arithmetic) mean of the reciprocals of the data values.

$$\text{harmonic mean} = \left(\frac{1}{N} \cdot \sum_{n=1}^{N} \frac{1}{x_n} \right)^{-1}$$

Having said that "average" should not be used, the verb form is difficult to avoid as there is no appropriate verb form of *mean*.

BALUN: **BAL**anced to **UN**balanced transformer. The word *balanced* means a differential system; there are two signals 180° phase shifted with respect to each other. The word *unbalanced* means that one end is grounded {earthed}. A dipole antenna, for example, needs to be driven from a balanced source. If only a single-ended signal is available, it needs to be put through a balun to provide a balanced drive signal. Depending on the frequencies involved, the balun could be an

[7] A. Francis, *Advanced Level Statistics* (UK: Stanley Thornes, 1979).

ordinary transformer, a ferrite core with turns on it, a ferrite bead with wires through it, or a short-circuited $\lambda/4$ sleeve over the coaxial feed wire [a *sleeve balun*].

A balun can either be used to convert from a single-ended signal to a differential signal, or vice versa.

BASIC INSULATION … is a term from electrical safety standards. Let's suppose there is some AC mains wiring resting on a metal plate. The metal plate is well earthed {grounded; connected to the protective conductor} so if the wire's insulation breaks down there is no risk of electric shock. The insulation in this case need only meet the *basic insulation* requirement. This level of insulation is only designed to keep the equipment operating. If the metal plate was not earthed, and in fact was accessible to a user, a higher standard of insulation would be required. This would then be classified as either *reinforced insulation* or *double insulation*.

Reinforced insulation is thicker than basic insulation and therefore able to withstand a greater applied voltage. Double insulation is two distinct levels of insulation, usually of different materials, each of which individually could withstand the requirements of the basic insulation tests.

BEAT FREQUENCY OSCILLATOR (BFO): Two relatively high frequency oscillators are combined in an RF mixer to extract the difference frequency, the *beat frequency*. One use is to create a swept frequency with a large high to low frequency range. It is difficult to make a basic oscillator variable over a 100:1 frequency range. However if an oscillator is made to vary over a 2:1 range and then the "offset frequency" is subtracted, a wide range results. For example, a 1 GHz oscillator mixed with a 1 GHz – 2 GHz variable oscillator produces a difference frequency from DC to 1 GHz.

BER … stands for Bit Error Ratio [or Bit Error Rate]. On a digital communication channel, the noise may flip a 0 to a 1, or vice versa. This is a 'bit error'. A BER of 10^{-10} means 1 error in 10^{10} bits. This may sound like very few errors, but with 100 MHz clocks on 16 bit data this means 1 error every 6 seconds.

BER & BEI … are the real and imaginary parts, respectively, of a *Bessel function* with a specific complex argument, known as the *Kelvin-Bessel functions*. (Ber = **Be**ssel **r**eal)

$$J_0\left(\frac{x}{\sqrt{j}}\right) \equiv ber(x) + j \cdot bei(x)$$

$$ber(x) = 1 - \frac{x^4}{2^2 \cdot 4^2} + \frac{x^8}{2^2 \cdot 4^2 \cdot 6^2 \cdot 8^2} - \ldots = 1 - \left(\frac{x^2}{4}\right)^2 \times \frac{1}{(2!)^2} + \left(\frac{x^2}{4}\right)^4 \times \frac{1}{(4!)^2} - \ldots$$

$$bei(x) = \frac{x^2}{2^2} - \frac{x^6}{2^2 \cdot 4^2 \cdot 6^2} + \frac{x^{10}}{2^2 \cdot 4^2 \cdot 6^2 \cdot 8^2 \cdot 10^2} - \ldots = \frac{x^2}{4} - \left(\frac{x^2}{4}\right)^3 \times \frac{1}{(3!)^2} + \left(\frac{x^2}{4}\right)^5 \times \frac{1}{(5!)^2} - \ldots$$

A Bessel function of an imaginary argument is known as a *modified Bessel function of the first kind.*

$$I_0(x) = J_0(j \cdot x)$$

The *ber* and *bei* functions occur in the solution to the ***skin effect*** problem in cylindrical conductors.

BESSEL FUNCTION … is a solution to *Bessel's equation*, a second order linear differential equation. Bessel's equation of order *n* is defined by …

$$\boxed{\frac{d^2y}{dx^2} + \frac{1}{x} \cdot \frac{dy}{dx} + \left(1 - \frac{n^2}{z^2}\right) y = 0}$$

Bessel's function of zero order, $J_0(x)$, is a solution to the above equation for *n*=0.

$$y = J_0(x) = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2 \cdot 4^2} - \frac{x^6}{2^2 \cdot 4^2 \cdot 6^2} + \ldots$$

Bessel functions naturally occur as solutions to electromagnetic problems involving cylindrical symmetry, hence they are sometimes referred to as *cylinder functions*. The general solution to

Bessel's equation of zero order, $\frac{d^2y}{dx^2}+\frac{1}{x}\cdot\frac{dy}{dx}+k^2y=0$, is:

$y=A\cdot J_0(kx)+B\cdot Y_0(kx)$, where *A* and *B* are constants required to satisfy the boundary conditions, *k* is a real constant {not imaginary or complex}, and $Y_0(\;)$ is a Bessel function of the second kind. The notation, name and value of Bessel functions of the second kind are remarkably inconsistent between authors. The function can also be called a *Neumann function* and can be written as $N_0(\;)$. Since $Y_0(0+)\to-\infty$, boundary conditions often mean that the constant *B* is zero in the general solution so that the Neumann function is not required.

For the differential equation $\frac{d^2y}{dx^2}+\frac{1}{x}\cdot\frac{dy}{dx}-k^2y=0$, the general solution is $y=A\cdot I_0(kx)+B\cdot K_0(kx)$, where $I_0(\;)$ is the *modified Bessel function of the first kind* and $K_0(\;)$ is a *modified Bessel function of the second kind*. Again it is found that $K_0(0+)\to\infty$.

For an angle-modulated signal, $V_O=\sin(\omega_c t+\theta\cdot\sin[\omega_m t])$, the carrier amplitude is $J_0(\theta)$ and amplitude of the first sideband pair (offset $\pm\omega_m$ from the carrier) is $J_1(\theta)$. The first sidebands and carrier are equal when $J_0(\theta)=J_1(\theta)$, $\theta=1.4347$. When $\theta=2.4048$, $J_0(\theta)=0$, meaning that the carrier is completely suppressed (zero).

BESSEL-THOMPSON FILTER [8] … is a set of maximally flat group delay filters (linear phase) for use on time domain circuits. The rolloff in the frequency domain is not as fast as ***Butterworth***, ***Chebyshev*** and ***Elliptic*** filters, but those all overshoot in the time domain, whereas the Bessel-Thompson does not. It originated in 1949, whereas the Butterworth originated in 1930.

BIFILAR WINDING … is a winding technique used to reduce inductance. Wrap a piece of wire many times around a bobbin and you have made an inductor. Take the same length of wire folded back on itself, making the *go* and *return* paths next to each other, and there is very little 'loop area' between the wires. Winding this doubled wire around a bobbin gives very little increase in inductance. However, this method greatly increases the capacitance between the ends of the wire. Regions of large potential difference between conductors give large amounts of self-capacitance.

For a low value resistor (say <10 Ω) the self-inductance will limit the high frequency response. For a higher value of resistor (say >1 kΩ) the self-capacitance will limit the high frequency response. Thus bifilar winding is only appropriate on low value resistors.

A better bifilar technique exists, however. Consider winding a wire as a clockwise helix around a cylindrical former. If a second wire is now wound over the top of the first wire, but as an anticlockwise helix, the magnetic fields will tend to cancel when the two wires are connected in parallel. Furthermore, since the wires will be at the same potential where they cross, the self-capacitance will be minimised. Using a flat former rather than a cylindrical former further improves the results and is known as the *Ayrton-Perry* construction.

BIOT-SAVART LAW: The discovery by Oersted (also written as Örsted) in 1820 that a current-carrying conductor influenced a magnetic compass needle opened up a whole new field of exploration. Within the next few weeks the basis of the modern theories was discovered and published,[9] Ampère, Biot & Savart being the key explorers.

Biot and Savart did their first experimental work using a magnetic compass needle and a long straight wire. A long bar magnet was used to cancel out the effect of the earth's magnetic field. The actual measurement of the magnetic force was done using the period of oscillation of a small

[8] W.E. Thomson, 'Delay Networks Having Maximally Flat Frequency Characteristics', in *Proceedings of the IEE*, 96 part 3 (1949), pp. 487-490.

[9] R.A.R. Tricker, *Early Electrodynamics, The First Law of Circulation* (Pergamon, 1965).

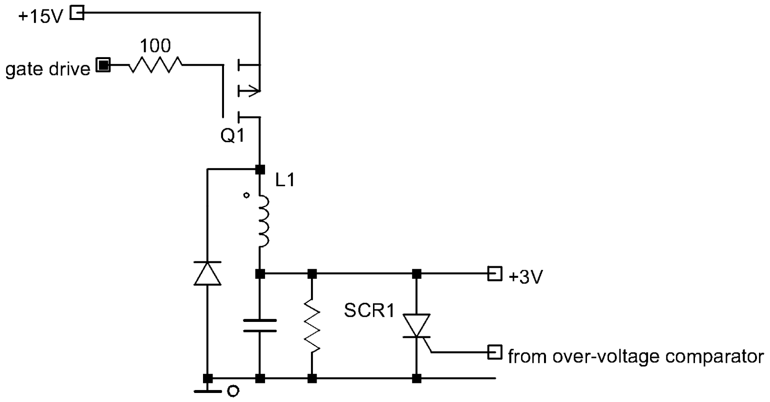
magnetic compass needle. When the needle was intentionally disturbed from its equilibrium position, the magnetic field provided the restoring force. This force then set the oscillation period of the needle, much as a pendulum's period is governed by the gravitational force.

BODE PLOT … is a plot of the ***loop-gain*** of a control system or an amplifier. The simplified intersecting straight-line form is known as an ***asymptotic*** Bode plot. The horizontal scale is of log[frequency] and the vertical scales are loop-gain (in dB) and loop phase. If the gain curve crosses the 0 dB axis at a rate of 40 dB/decade, the system will either be unstable or marginally stable.

"BRICKWALL" FILTER … is a filter response with infinite attenuation outside of the pass-band. It is a physically unrealisable theoretical device. A simple single-pole RC filter rolls off gradually, having a slope of 20 dB attenuation increase for every decade (10×) increase in frequency. The concept can be used, for example, to work out an equivalent noise bandwidth, allowing RMS noise to be readily calculated. A brickwall filter has a ***shape factor*** of 1, the best shape factor that can be achieved. See also the diagram under *shape factor*.

BROWN-OUT: An 'outage' is a period during which a service or supply has been stopped; in this case it is the AC (mains) supply being considered. A *black-out* is where all power supply has ceased, not least of which is to lighting circuits. A *brown-out* is less "black and white". If the AC supply drops below its amplitude spec then that is a brown-out. Under these circumstances the operation of power supply circuits becomes uncertain. It is important that any PSU should be able to deal with this situation in a defined and non-destructive manner. The PSU doesn't necessarily have to continue to supply its full rated output, but it must not start oscillating, *hunting* or smoking!

BUCK REGULATOR … is the one of the most efficient basic topologies for a switched-mode power supply because the inductor is not having to store all of the energy supplied each cycle.



If the series pass device, Q1, goes short-circuit then the 3 V output is going to try to rise to nearly 15 V. An over-voltage trip should be used to trigger the ***crowbar*** SCR. This will prevent circuitry powered from the +3 V rail from being destroyed. The crowbar would then blow a power fuse, not shown, in series with L1.

BUDGET: Engineers talk about "error budgets" and "power budgets". The English definition of budget is related to an available amount of money and how it will be spent. In the electronics world it is the available amount of the explicitly mentioned quantity, and how this quantity will be distributed or consumed. Hence in a "power budget" you might allocate 100 mW to signal conditioning, 50 mW to digital I/O, and 50 mW to the display device, for example. In the first instance of a new design you have to estimate how much power each section will consume, thereby enabling a power supply spec to be completed.

BURDEN VOLTAGE: Ammeters ordinarily have a volt drop across themselves when measuring current. This volt drop is called the *burden voltage*. This is a necessary feature of a passive meter such as a simple moving coil meter, but it is not an essential part of an active DMM circuit.

In a digital ammeter it is convenient and simple to use an internal resistor to generate a voltage and then measure the volt drop across it. Rather than specify burden voltage, it is equally useful to have the effective resistance of the ammeter specified.

Clamp-on ammeters do not have burden voltages because they do not interrupt the current path. DC-capable clamp-on ammeters would use either Hall Effect sensors or GMR (Giant Magneto-Resistive) sensors. AC only clamp meters use a split ferromagnetic core with a wound coil.

BURN-IN … is a process of running a component or system for an initial period to try to get weak components and interconnections to fail (*infant mortality*), rather than failing in service with a customer. It is more expensive to fix a fault in a customer's finished unit than it is to fix a sub-assembly or component. Burn-in is often done at elevated temperature and with the power being switched on and off repetitively. The temperature may also be cycled rapidly in order to encourage the weak components or solder joints to fail.

BUTTERWORTH FILTER [10] …is a class of filters with "maximally flat" ***monotonic*** frequency response which originated in 1930. For a Butterworth filter of order *n*, the magnitude of the transfer function is $|T| = \dfrac{1}{\sqrt{1+\left(\dfrac{f}{B}\right)^{2n}}}$, where *B* is the 3 dB bandwidth. The definition of "maximally flat" means that all derivatives of the transfer function are zero at DC. As the frequency gets closer to DC, the Butterworth filter will always be closer to the LF asymptotic gain than any other filter. However, if a lower error limit is specified (below which errors are neglected), it is always possible to find a transfer function which gives a sharper roll-off after this limiting point. The monotonic-L by Papoulis[11] is a specific class of filters which is characterised by a maximum slope at the 3 dB cutoff frequency. However, the frequency response, whilst still monotonic, has a nasty plateau in it. There have been numerous other successful efforts at "better than Butterworth" monotonic responses.[12] All of this can get a bit 'academic' because the tolerance of the filter components can also make the response not-monotonic.

Chebyshev and ***Elliptic*** filters roll off faster, but have amplitude ripples in the passband.

It is important to realise that cascading two 2-pole Butterworth filters does not give a 4-pole Butterworth filter; it will give a 4-pole filter, but the response will not be optimal and will not follow the Butterworth transfer function given above.

The first 7 Butterworth polynomials are:

$$B(1) = s+1$$
$$B(2) = s^2 + 1.41421s + 1$$
$$B(3) = s^3 + 2.00000s^2 + 2.00000s + 1$$
$$B(4) = s^4 + 2.61313s^3 + 3.41421s^2 + 2.61313s + 1$$
$$B(5) = s^5 + 3.23607s^4 + 5.23607s^3 + 5.23607s^2 + 3.23607s + 1$$
$$B(6) = s^6 + 3.86370s^5 + 7.46410s^4 + 9.14162s^3 + 7.46410s^2 + 3.86370s + 1$$
$$B(7) = s^7 + 4.49396s^6 + 10.09783s^5 + 14.59179s^4 + 14.59179s^3 + 10.09783s^2 + 4.49396s + 1$$

The even-order polynomials are conveniently factorised into two-pole sections for use in Sallen-Key filter stages:

$$B(4) = \left(s^2 + 0.765367s + 1\right)\left(s^2 + 1.84776s + 1\right)$$

$$\boxed{B(n) = \prod_{k=1}^{\frac{n}{2}} \left(s^2 + 2s \cdot \cos\left[\left(k - \frac{1}{2}\right) \cdot \frac{\pi}{n}\right] + 1 \right)}$$

The odd-order polynomials are similarly given by:

$$B(5) = (s+1)\left(s^2 + 0.618034s + 1\right)\left(s^2 + 1.618034s + 1\right)$$

$$\boxed{B(n) = (s+1) \cdot \prod_{k=1}^{\frac{n-1}{2}} \left(s^2 + 2s \cdot \cos\left[k \cdot \frac{\pi}{n}\right] + 1 \right)}$$

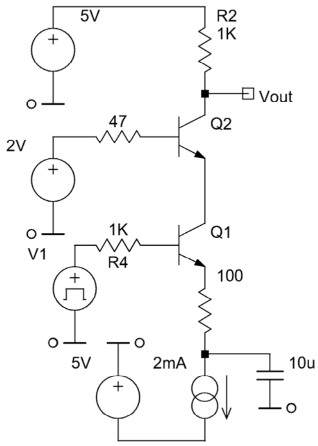
[10] S. Butterworth, 'On the Theory of Filter Amplifiers', in *Experimental Wireless & The Wireless Engineer* (Oct 1930), pp. 536-541.

[11] A. Papoulis, 'Optimum Filters with Monotonic Response', in *Proceedings of the Institute of Radio Engineers*, 46 (March 1958), pp. 606-609.

[12] B.D. Rakovich, and S.M. Lazovich, 'Monotonic Low-Pass Filters with Improved Stopband Performance', in *IEEE Transactions on Circuit Theory*, CT-19 (March 1972), pp. 218-221.

Note that for a second order stage, critical damping is achieved with the *s* coefficient set to 1.4142, as seen for the B(2) polynomial. When the *s* coefficient is smaller than this value the individual stage peaks in the frequency domain.

CASCODE CONNECTION:

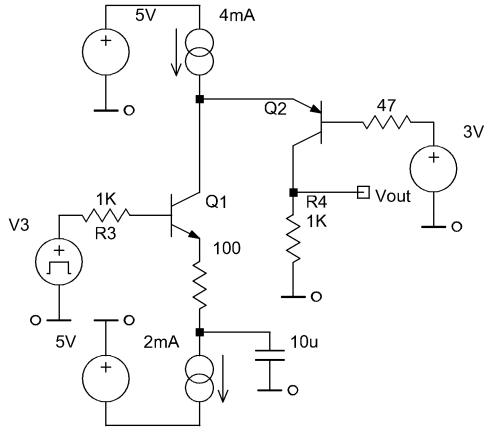


The input signal is V1, the source resistance is R1. Q1 amplifies the signal producing a current. If this current were fed directly to the load resistance R2, the bandwidth would be reduced because of the collector-base capacitance of Q1. This is called the ***Miller capacitance***, because the voltage gain makes the effective capacitance larger.

By feeding the signal into a common-base stage, very little signal voltage is developed on Q1 collector. This gives Q1 a low voltage gain, leading to less Miller capacitance and therefore higher bandwidth. In this simulation circuit, adding Q2 gives 4× the bandwidth.

In general, a cascode connection is a pair of active devices working together. The first device produces a current gain. The signal current is not used to generate a voltage because the Miller capacitance would limit the bandwidth. Instead the signal is routed into a low impedance to high impedance conversion stage. This isolates the feedback capacitance and increases the bandwidth.

In the previous scheme the signal current keeps going *up* {towards the positive power rail}. By using a complementary {opposite polarity; other 'sex'} transistor, the signal can be reflected back *down*. This scheme is known as a *folded cascode*. It is used to keep the signal within the confines of the power rails.



CAVITY MODES: Cavities can be used as high-Q structures for oscillators, but can also limit the frequency of operation of amplifiers due to parasitic oscillation at one of the cavity modes. In a simple air-filled conducting rectilinear cavity, the modes are expressed by three integer mode numbers, only one of which can be zero. Numbering from the longest side, to the next longest, and finally to the shortest, TE_{110} is the lowest cavity mode.

$$f_{m,n,p} = c \times \sqrt{\left(\frac{m}{2a}\right)^2 + \left(\frac{n}{2b}\right)^2 + \left(\frac{p}{2d}\right)^2}$$

c is the speed of light: *a*, *b* and *d* are the length, width and depth respectively. The formula applies equally to the TE_{mnp} and the TM_{mnp} modes.

For use in oscillators it is important to operate the cavity in only one mode, with a good separation to the next higher mode. A good separation is achieved by making *a*=*b*, with *d* much smaller, and running in the TE_{110} mode. The next higher modes, TE_{120} and TE_{210}, are then a factor of ×1.58 higher in frequency.

A slightly better separation can be achieved by running in TE_{110} with *b* = 0.61*a*, resulting in a ×1.78 separation between the next higher modes TE_{120} and TE_{310}. The TE_{210} mode has to be avoided by putting both the stimulus and extraction probes into the cavity anywhere along the centreline of the long side.

CAUER-PARAMETER FILTER: Also known as an elliptic filter [because the poles lie roughly on an ellipse]. This family of filters has ripples in the passband and the stop band, but achieves a steeper rolloff than the ***Bessel***, ***Butterworth*** and ***Chebyshev*** filters [of the same order]. It is unsuitable for time domain signal processing due to excessive overshoot (15%).

CCD: **C**harge **C**oupled **D**evice. There are two types of CCD in common use. One is an optical imaging CCD and the other is a signal sampling CCD. In the signal sampling CCD, charge is put into the device through an injector structure. It is then shifted through the array by a multi-phased clock system. In the optical imaging CCD, charge is produced on internal capacitors by photo-emission. This charge represents the image. The charge is then read out of the CCD by the usual CCD charge transfer mechanisms.

CENTRAL LIMIT THEOREM: If many different variables with unspecified random distributions are added together, the resulting distribution will tend towards a Gaussian (Normal) distribution. The proof of this theorem is remarkably complicated.[13]

It is found in nature that complex systems tend to have Gaussian distributions of noise, length, weight &c.

CERMET: A generic material used in the construction of resistors and potentiometers, made from a mix of **CER**amic and **MET**al which is screened onto a substrate and baked, a *thick-film* process. There is no specific composition for cermet because every manufacturer has their own proprietary [secret] mix, but the metals used include silver, palladium, platinum, ruthenium, rhodium and gold. It gives much better stability than carbon and it has a TC which is 10× lower.

CHEBYSHEV FILTER: Also spelt Tschebyscheff. A group of filters with defined ripple in the passband, but a monotonic rolloff. For a given order of filter [number of poles] the Chebyshev rolls off faster than both the ***Bessel*** and the ***Butterworth***. Faster rolloff is achieved by allowing more ripple in the passband. There are therefore two parameters to choose when using a Chebyshev filter, the number of poles and the amount of ripple in the passband. This filter family is not suitable for time domain signal processing due to the large overshoot and ringing, these being even worse than those achievable using a Butterworth response.

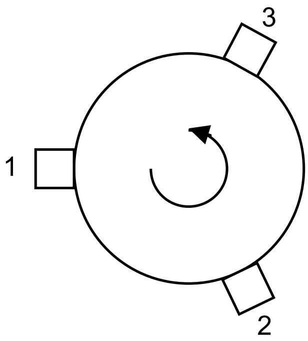
CHEBYSHEV POLYNOMIALS: Also spelt Tschebyscheff, and other variants, because of the translation from Russian.

These ***orthogonal*** polynomials are solutions to what otherwise appear to be insoluble trigonometric problems.

First Kind	Second kind
$T_n(x) \equiv \cos(n \cdot \operatorname{arc}\cos(x))$	$U_n(x) \equiv \sin(n \cdot \arccos(x))$
$T_1(x) = x$	$U_1(x) = \sqrt{1-x^2}$
$T_2(x) = 2x^2 - 1$	$U_2(x) = 2x\sqrt{1-x^2}$
$T_3(x) = 4x^3 - 3x$	$U_3(x) = (4x^2-1)\sqrt{1-x^2}$
$T_{n+1}(x) = 2x \cdot T_n(x) - T_{n-1}(x)$	$U_{n+1}(x) = 2x \cdot U_n(x) - U_{n-1}(x)$

[13] J.V. Uspensky, 'Ch XIV: Fundamental Limit Theorems', in *Introduction to Mathematical Probability* (USA: McGraw-Hill, 1937), pp. 283-307.

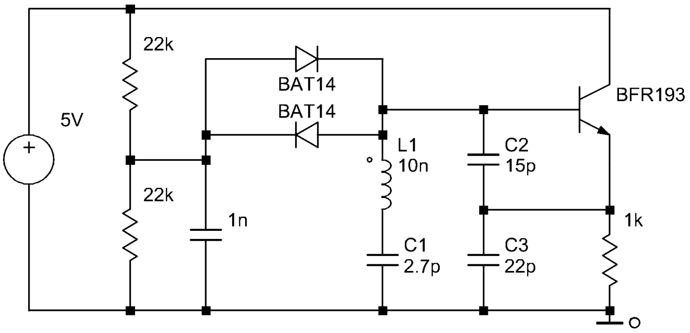
CIRCULATOR:



A highly non-linear, non-reciprocal device which is markedly directional, and available for use above 100 MHz. In this 3-port device, power flows easily from ports 1 to port 2, from port 2 to port 3 and from port 3 to port 1. By easily is meant the insertion loss is low, <1 dB. In the reverse directions, port 2 to port 1, port 1 to port 3 and port 3 to port 2, the insertion loss is high, >20 dB.

If one of the ports is Z_0-terminated you have a two-port device known as an ***isolator***; the RF/microwave equivalent of a mechanical one-way valve.

CLAPP OSCILLATOR: This oscillator was designed to minimise the effects of changes in the oscillation frequency due to changes in the amplifying device, and also to make band-switching easier. The original paper quotes a frequency stability of better than 1ppm for ±15% supply variation to the thermionic valve oscillator.[14]



This circuit shows the basic Clapp configuration, consisting of L1 and the series combination of C1, C2 and C3. This circuit simulation oscillates at approximately 1 GHz. Note that C2 and C3 are much larger than C1, allowing C1 to dominate the setting of the resonant frequency.

This configuration is sometimes known as a series-tuned ***Colpitts*** oscillator. The difference between the ***Clapp*** and the Colpitts is the capacitor in series with the inductor. This enhances the Q and allows simple switching of the frequency by the selection of a new LC path.

CLASSICAL THEORY… really means a non-quantum theory, but it can also mean a theory that neglects *special relativity*. Thus one tends to think of classical theories as being prior to 1900. In the era of 1900 to 1930, physics took a whole new direction and therefore 1900 is a rough demarcation point between the 'old' classical physics and the new quantum-relativistic physics.[15] In many respects, classical physics is easier to understand and to teach because the concepts are easy to visualise in terms of models. In quantum physics, attempts to think in terms of models always seem to end up giving misleading and unrepresentative results.

CLASS I … equipment is earthed {grounded} and uses the earth {protective conductor} as a safety mechanism against electric shock. ***Creepage and clearance*** distances are not required to be too large because fault currents are shunted {shorted} safely to earth.

Class II equipment uses double insulation or reinforced insulation. You see this on items like domestic power tools and low-power 'battery eliminators'. The symbol with one square completely inside another is the *double insulation* symbol. There is no earth needed or used. The creepage and clearance distances are made larger than the Class I requirements to cope with this fact. In fact the distances and test voltages are usually double those of Class I insulation.

Another completely separate definition of Class I is as the dielectric for capacitors. Class I is for the temperature compensated NP0 type of dielectrics.

[14] J.K. Clapp, 'An Inductance-Capacitance Oscillator of Unusual Frequency Stability', in *Proceedings of the Institute of Radio Engineers* (March 1948), pp. 356-358.

[15] G. Gamow, *Thirty Years That Shook Physics: The Story of Quantum Theory* (Dover Publications, 1966).

CLOSED-FORM SOLUTION … is defined as one which gives the dependant variable *explicitly* in terms of one or more independent variables; a result given recursively, iteratively, or implicitly, is not in a *closed-form*.

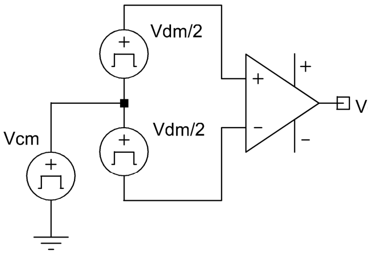
There is some debate as to whether or not a power series solution constitutes a closed-form solution. It is generally agree that the elementary functions are acceptable in a closed-form solution, but more complicated functions such as *hyper-geometric functions* and *Bessel functions* are more contentious.

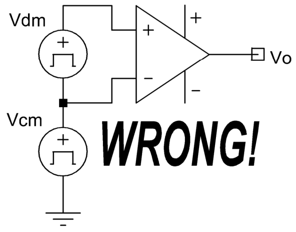
Given the ready availability of computing power, I would argue that any existing tabulated or graphed function should be acceptable in a closed-form solution. The key question is whether or not somebody reading the solution to the problem can plot the result. If the function used is not a standard library function, but is a recognised named function that can be readily evaluated without numerical integration, then it is reasonable to call that a closed-form solution.

CMRR: **C**ommon-**M**ode **R**ejection **R**atio. The text book definition is:

$$\boxed{CMRR = \frac{\text{Differential - Mode gain}}{\text{Common - Mode gain}}}$$

… which is usually expressed in dB by taking 20·log$_{10}$ ( ) of the above ratio. CMRR is a good thing and you want lots of it. The common-mode voltage is not the signal of interest; it is an *interfering voltage* which needs to be reduced.





The common-mode signal is defined as the (instantaneous) mean of the two input signals. The differential-mode signal is defined as the (instantaneous) difference between the two input signals. You will see that the equivalent circuit presented above on the left correctly models these definitions; the circuit underneath is incorrect. Nevertheless you will still see this incorrect equivalent circuit widely presented in books, papers and application notes.

The reason why the incorrect equivalent circuit persists is that for some applications the difference between the two models is quite small. Consider an amplifier with a (differential mode) gain of several hundred. When the signal is in the region of 10 mV and the common-mode voltage is in the region of several volts, the error in the incorrect model is usually small enough to neglect.

If one side of a differential amplifier is grounded, and a signal is applied to the other side, a common-mode signal is still being applied; the common-mode signal is half the input signal.

For a single-ended output differential amplifier, both the common-mode input signal, V_{CM}, and the differential-mode input signal, V_{DM}, combine to produce the single output voltage V_O:

$$V_O = G_{DM} \cdot V_{DM} + G_{CM} \cdot V_{CM}$$

Where *G* is the voltage gain, with self-evident subscripts to specify the mode.

The effect of the common-mode signal at the output is not distinguishable from the effect of the differential-mode signal, there being only one output pin. The effect of the common-mode signal at the output could therefore be attributed to an extra differential-mode signal at the input. Let's call this new input signal "the output common-mode signal referred to the input", V_X. By definition then: $G_{CM} \cdot V_{CM} = G_{DM} \cdot V_X$

In effect, a common-mode signal at the input to a differential amplifier produces an *equivalent* input signal of size $V_X = V_{CM} \cdot \frac{G_{CM}}{G_{DM}}$ and therefore $\boxed{CMRR = \frac{\text{Differential Mode gain}}{\text{Common Mode gain}} = \frac{G_{DM}}{G_{CM}} = \frac{V_{CM}}{V_X}}$

This is a very important result. At DC, for example, the input offset voltage of an opamp changes with common-mode (input) voltage. A voltage follower with a CMRR of 80 dB will change its input offset voltage in response to a change of common-mode signal of 5 V by the amount of

$$V_X = \frac{V_{CM}}{CMRR} = \frac{5}{10^{80/20}} = 0.5\,\text{mV}$$

This idea is essential in correctly applying information from opamp data sheets. It also allows you to calculate how much CMRR is necessary to make a small signal measurable in the presence of high common-mode interference.

If the signal has a repeating pattern, and a synchronous signal is available which is not itself corrupted by common-mode noise, this synchronous signal can be used as a trigger for a digital storage oscilloscope. By averaging the signal, the common-mode rejection, if poor, can be considerably improved (say >20 dB). This technique presupposes that the differential-mode signal is entirely asynchronous to the common-mode signal.

COAXIAL TRANSMISSION LINES:

The *characteristic impedance* of a circular coaxial transmission line (coax) is:

$$\boxed{Z_0 = \frac{60}{\sqrt{\varepsilon_r}} \cdot \ln\left(\frac{\text{inner radius of outer conductor}}{\text{outer radius of inner conductor}}\right)} = \frac{138.16}{\sqrt{\varepsilon_r}} \cdot \log_{10}\left(\frac{\text{inner radius of outer conductor}}{\text{outer radius of inner conductor}}\right)$$

where ε_r is *relatively permittivity* of the insulator, also known as its *dielectric constant*.

The characteristic impedance for loss-less coax is a pure resistance, $Z_0 = \sqrt{\frac{L}{C}}$,

L and *C* being the inductance and capacitance per unit length of line respectively.

When the losses are finite, $Z_0 = \sqrt{\frac{L}{C}} \cdot \sqrt{\frac{1+R/j\omega L}{1+G/j\omega C}}$, which contains a reactive part.[16]

Below low audio frequencies the characteristic impedance approaches $Z_0 = \sqrt{\frac{R}{G}}$.

R and G are the resistance and conductance per unit length of line respectively.

Above audio frequencies, and when the losses are small, $\boxed{Z_0 \approx \sqrt{\frac{L}{C}} \cdot \left[1 - \frac{j}{2}\left(\frac{R}{\omega L} - \frac{G}{\omega C}\right)\right]}$

The attenuation constant is $\alpha = \sqrt{RG}$ ***Nepers*** per unit length at sub-audio frequencies.

This rises to $\alpha = \frac{1}{2}\left(\frac{R}{Z_0} + G \cdot Z_0\right)$ Nepers per unit length at radio frequencies.

The RF attenuation is therefore $4.343\left(\frac{R}{Z_0} + G \cdot Z_0\right)$ dB per unit length.

For transmission lines in general, the line ‘constants’ R, L, C and G are far from constant. $R \propto \sqrt{f}$ due to the skin effect. $G \propto f$ due to dielectric loss. These effects cause fast transients (such as lightning strikes) on power lines to dissipate much more rapidly than would be predicted if the line

[16] S.Y. Liao, 'Chapter 2-1-1: Transmission-Line Equations and Solutions' in *Microwave Circuit Analysis and Amplifier Design* (Prentice-Hall International, 1987), pp. 10-15.

constants were fixed.[17]

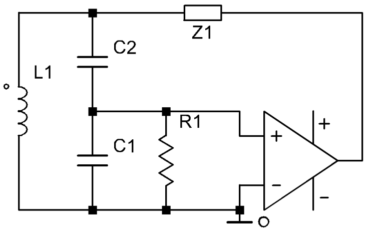
50 Ω coaxial cable comes in many different types, specified by overall diameter, dielectric material, screen construction and signal conductor support structure. All of them have a tolerance on this 50 Ω spec, of course. Very thin coax is the worst on tolerance as well as on loss; a tolerance of ±10% on the characteristic impedance is not unusual.

It is a matter of practical experience that coaxial cables can interact with each other above around 300 MHz. The screens become less effective so that cross-talk and UHF oscillations are possible. It may therefore be necessary to physically separate coaxial cables in analog circuits operating at these frequencies. The separation only needs to be 1 mm, but this can be enough to prevent unintentional coupling. Alternatively, rigid or semi-rigid coax assemblies may be used.

The transport mode in coax cable is TEM, *transverse electromagnetic*. All the preceding theory is based on TEM waves in the coax. If a cable is made too large for the operating frequency then a $TE_{1,1}$ mode can exist. This is very bad because the TEM wave and the TE wave travel at different *group velocities* causing dispersion of pulses, and mismatches. This problem is referred to as *over-moding*.

A first approximation to the critical wavelength is $\lambda_{1,1} \approx \frac{\pi}{\sqrt{\varepsilon_r}} \cdot (R + r)$, where *R* is the inner radius of the outer conductor, and *r* is the outer radius of the inner conductor. 50 Ω cable for use up to 5 GHz using PTFE dielectric must therefore be < 40 mm in diameter, and for use up to 50 GHz must be < 4 mm in diameter, where the diameter is measured to the inside of the outer conductor.

COLPITTS OSCILLATOR … is a sinusoidal LC oscillator with a 'tapped' capacitor chain which originated in 1918.[18]



In this generic representation, the resonant circuit is formed by the series capacitance of C1 and C2, in parallel with the inductance of L1. R1 represents the finite input resistance of the amplifier and Z1 is a high impedance which includes the losses in L1 and C2. A voltage output amplifier would ordinarily drive the resonant circuit ("tank") via a loose reactive coupling, a small coupling capacitor being one example.

It is usual for C1 to be larger than C2 so the input impedance of the amplifier does not severely limit the resonant circuit Q. Also, variation in the input capacitance of the amplifier does not change the operating frequency to such a large degree when C1 is large.

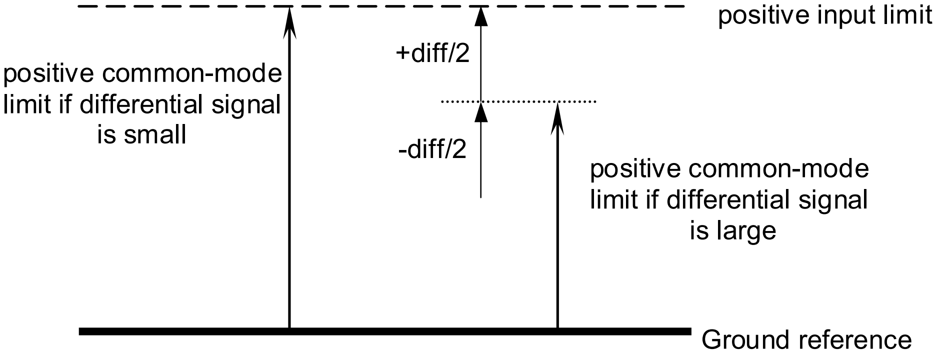
The output of the amplifier (before Z1) is not accessible when the amplifier has a current output, this circuit then being the Thévenin equivalent. Rather than take the output from the high impedance tank circuit, thereby ruining the Q, it is usual to take the overall output from the bottom of the capacitor chain, that is the *input* to the amplifier. This point is made a low impedance by making C1»C2.

If the inductor is tapped instead of the capacitor then that would be a ***Hartley oscillator***.

COMMON-MODE INPUT RANGE (CMIR) … is the range of input voltage over which a differential amplifier is specified for linear operation. The common-mode rejection ratio (***CMRR***) will be given over a specific common-mode input range. If this range is exceeded then the CMRR is no longer guaranteed. In fact it is likely that the amplifier will start limiting and the CMRR will degrade very rapidly. The actual voltage at which clipping occurs will be a limiting value for either terminal. It is up to the manufacturer how they specify this. CMIR is also called *common-mode voltage range*, CMVR.

[17] C.P. Steinmetz, 'The General Equations of the Electric Circuit III; Variation of Constants r, L, C, and g, and Its Effects.', in *Proceedings of AIEE*, 38 (Feb 1919), pp. 249-318.
[18] E.H. Colpitts, 'Improvements Relating to Signalling by High-Frequency Currents, as in Wireless Telegraphy', *UK Patent Specification 141,047* (USPO, 1914: UKPO, 1921).

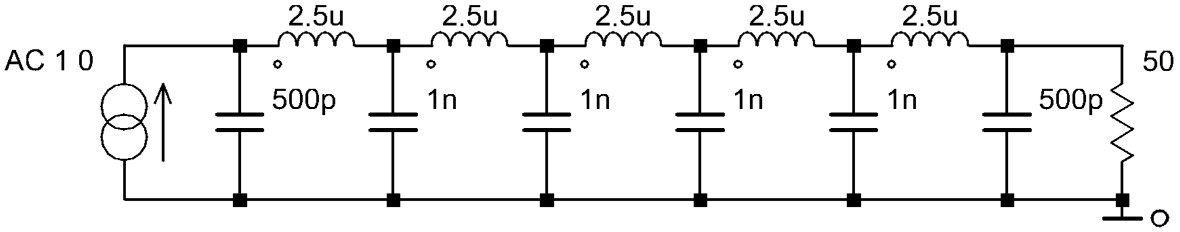
The best spec [biggest numbers, looks good on a data sheet] is achieved with a small differential signal 'on top of' a large common-mode signal. The common-mode input range can then be right up to the input clipping limits. Otherwise the common-mode input range would have to be specified as this previous limit minus half the allowed differential signal.


positive input limit
positive common-mode limit if differential signal is small
+diff/2
-diff/2
positive common-mode limit if differential signal is large
Ground reference

You may see CMIR written as CMR. This is confusing as CMR is defined as "common-mode rejection expressed in dB" by some people. It is therefore best not to use the abbreviation CMR for any purpose.

CONSTANT-K FILTER ... is method of filter design originated by Zobel in 1923 and is defined in his own words:

"The 'constant *k*' wave-filter belonging to any class is defined as that ladder type wave-filter whose product of series and shunt impedances, and therefore characteristic impedance, *k*, of the corresponding smooth line, is constant independent of frequency." [19]



This is a lumped approximation of a lossless 50 Ω transmission line which follows the rules given above for a constant-*k* filter. Notice that the shunt elements at the beginning and end of the line have been halved. Alternatively the line could have been ended using inductors of half the usual values.

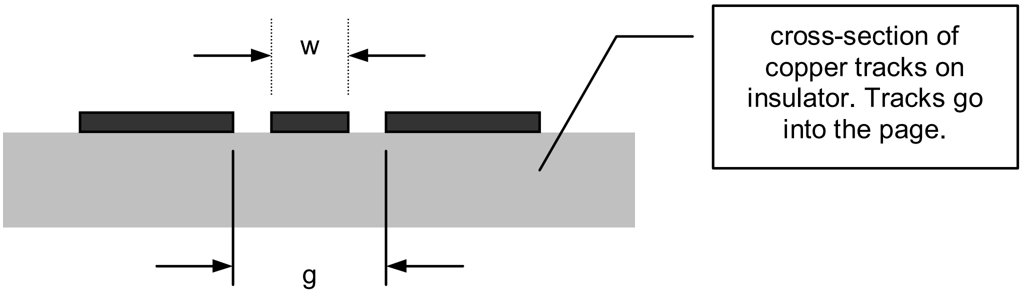
Do not use this model as a lumped equivalent of a transmission line above $f = \frac{1}{20\sqrt{LC}}$, since a lossless transmission line does not filter the signal.

The constant-*k* method was the first major treatment of iterated filter structures. It was superseded by ***m-derived filter*** theory and the whole gamut of modern filter techniques. The constant-*k* method avoided higher mathematics, but the input impedance, flatness, and attenuation characteristics were poor by modern standards. The professional design procedure is to use a set of filter tables (or software) to get optimised values of the reactances, thereby avoiding the mathematics.

COPLANAR WAVEGUIDE: [CPW] ... is simply conductors on an insulator, where the signal conductor has ground conductors on both sides of it; there is no ground plane underneath. Controlled impedance lines on PCBs can be done in this way, instead of the more usual ***microstrip*** method. The ground strips need to be connected together with low impedance straps every now and then to prevent the build up of higher frequency modes. As a guideline, the ground strips would be between 3 and 10 times as wide as the signal track. This method of interconnecting microwave

[19] O.J. Zobel, 'Theory and Design of Uniform and Composite Electric Wave-Filters', in *Bell System Technical Journal*, 2, no. 1 (Jan 1923), pp. 1-46.

circuits originated in 1969.[20]


w
g
cross-section of copper tracks on insulator. Tracks go into the page.

The field pattern is complicated, but can be solved for an ideal case by the use of *conformal mapping*. The assumptions for the ideal case are that the ground strips are infinitely wide, the dielectric is infinitely thick and that the conductors have no thickness. The resulting formulae for the capacitance and impedance contain ***elliptic integrals***. These integrals, which are analytically insoluble in terms of elementary functions, have traditionally been tabulated.[21]

As shown in the diagram above, define *w* as the centre strip width and *g* as the total gap between the inner edges of the ground plane. The capacitance (per unit length) from the centre strip to ground is approximated by:

$$\boxed{C = \frac{\pi\varepsilon_0(\varepsilon_r + 1)}{\ln\left(\frac{4g}{w}\right) - 0.324 \times \left(\frac{w}{g}\right)^{2.2}}}$$

for $0.0 < \frac{w}{g} \le 0.5$ [*<0.03% calculation error*]

$$\boxed{C = \frac{2\varepsilon_0(\varepsilon_r + 1)}{\pi} \cdot \ln\left(\frac{1.999 - 2\sqrt{w/g}}{1 - \sqrt{w/g}}\right)}$$

for $0.4 \le \frac{w}{g} < 1.0$ [*<0.01% calculation error*]

The "calculation errors" given are relative to the approximations stated above, they are not absolute errors. The errors involved in the approximations used to create the elliptic integral solution could be a few percent and the tolerance on the dielectric constant could be ±10%.

Notice that for zero thickness conductors, the field pattern above and below the conductors is the same, regardless of the presence of the dielectric. The effective dielectric constant is therefore the mean of the two dielectric constants, giving $\boxed{\varepsilon_{eff} = \frac{\varepsilon_r + 1}{2}}$

To get the characteristic impedance of the coplanar waveguide you could now calculate the inductance and use $Z_0 = \sqrt{L/C}$. However, it is easier to use the fact that the velocity on a transmission line is $v = \frac{1}{\sqrt{LC}}$, which gives $Z_0 = \frac{1}{vC}$. But the velocity is dependant only on the effective dielectric constant, $v = \frac{c}{\sqrt{\varepsilon_{eff}}}$, where *c* is the speed of light.

$$\boxed{Z_0 = \frac{84.79}{\sqrt{\varepsilon_r + 1}} \cdot \left[\ln\left(\frac{4g}{w}\right) - 0.324 \times \left(\frac{w}{g}\right)^{2.2}\right]}$$

for $0.0 < \frac{w}{g} \le 0.5$

[20] C.P. Wen, 'Coplanar Waveguide: A Surface Strip Transmission Line.', in *IEEE Transactions on Microwave Theory and Techniques*, MTT-17, no. 12 (Dec 1969), pp. 1087-1090.

[21] M. Abramowitz, and I. Stegun, 'Chapter 17, Elliptic Integrals', in *Handbook of Mathematical Functions* (National Bureau of Standards, 1964; repr. Dover Publications, 1965).

$$\boxed{Z_0 = \frac{418.4}{\sqrt{\varepsilon_r + 1}} \cdot \frac{1}{\ln\left(\dfrac{1.999 + 2\sqrt{\dfrac{w}{g}}}{1 - \sqrt{\dfrac{w}{g}}}\right)}}$$

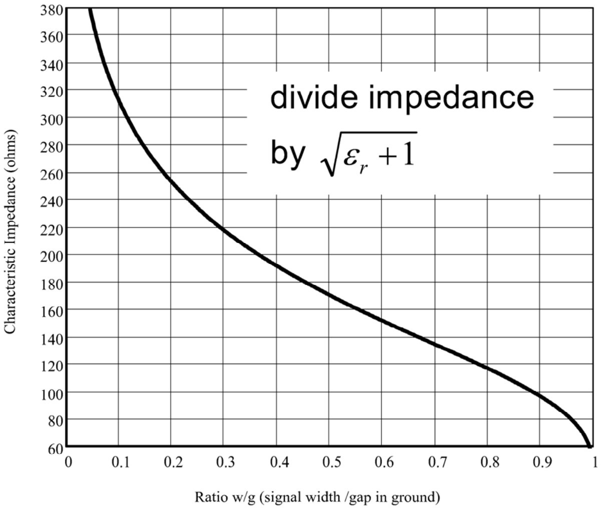
for $0.4 \le \dfrac{w}{g} < 1.0$

To find the actual characteristic impedance, divide the reading taken from the graph by the dielectric term.

Consider the case of a centre conductor of width *W*, with a gap of *W* on either side. In this case the ratio *w/g* is 1/3. The reading from the graph is 208 Ω. If the substrate has a dielectric constant of 4, then divide the 208 by $\sqrt{5}$ giving 93 Ω.

For a 50 Ω line in material with a dielectric constant of 4, w/g=0.8272.

The dielectric thickness does not affect the characteristic impedance by more than a few percent provided that it is at least twice as thick as the ground gap, *g*.



To achieve a given value of Z_0 arithmetically, inverse formulae are desirable.

$$\boxed{\frac{w}{g} = \left(\frac{\exp\left(\dfrac{418.4}{Z_0\sqrt{\varepsilon_r + 1}}\right) - 1.9985}{\exp\left(\dfrac{418.4}{Z_0\sqrt{\varepsilon_r + 1}}\right) + 2.000}\right)^2}$$

$Z_0\sqrt{\varepsilon_r + 1} \le 200\,\Omega$ [*<0.01% calculation error*]

$$\boxed{\frac{w}{g} = \frac{4}{\exp\left(\dfrac{Z_0\sqrt{\varepsilon_r + 1}}{84.79}\right) + \dfrac{164}{\left(Z_0\sqrt{\varepsilon_r + 1}\right) - 5} - 0.458}}$$

$Z_0\sqrt{\varepsilon_r + 1} \ge 180\,\Omega$ [*<0.02% calculation error*]

CREEPAGE AND CLEARANCE … relate to safety standards and high voltages. *Creepage* is the shortest distance across an insulating surface between two conductors. It is important that distances be kept large enough to prevent arcing (tracking) across the surface of the insulator. It is necessary to use a larger distance when the surface can be contaminated by dust and other material. If the surface is not flat, the standards allow the creepage distance to be the path length along the surface, provided that the distances along each surface are larger than say 2 mm. If the surface has fine (1.5 mm) grooves, these will not be sufficient to be considered as increasing the creepage distance over the 'line of sight' distance if there is substantial surface contamination. (Read the standards for the detail.)

When dealing with mains circuits at 230 V AC, typical creepage distances would be in the region of 3 mm to 6 mm, but it is essential to first decide whether the creepage distance required is for ***basic insulation*** purposes or *reinforced insulation*. A creepage distance of less than 1000 V/mm is never acceptable for any purpose.

Clearance is the shortest distance through air between two conductors. Clearance distances need to be increased with altitude to allow for the increased ease of arcing.

When you need to deal with Creepage and Clearance distances for safety purposes, read the appropriate standards carefully. EN61010-1:2001, Annex C provides useful diagrams, as does EN60950-1:2006 Annex F.

CREST FACTOR … is defined as the ratio of peak voltage to RMS voltage. It is particularly important in ***true RMS*** DVMs because the internal circuits can get overloaded without the meter reading over-range. This internal overload could give a grossly incorrect reading and there would be nothing to warn the user that an overload was happening.

The term was first suggested by G.Kapp, earlier than 1910.

CROSS-TALK … is unwanted coupling from one channel (or circuit) to another. The term originates from early telegraph and telephone systems. The use of a common earth wire in telegraph systems was eliminated, circa 1892, to minimise cross-talk. In 1895 Prof. Hughes introduced the idea of twisted pairs to minimise cross-talk on telephone systems.

CROWBAR … is a protection circuit used to shut down a power supply before damage can be done by an over-voltage fault. In a series regulator situation, the semiconductor can become short-circuit. The resulting over-voltage may damage components supplied by this power rail. To prevent damage, an over-voltage detection circuit puts a short-circuit straight across the regulated power rail. It is as though somebody got a crowbar {a heavy steel bar used to lever things apart} and dropped it across the power rails.

In practice the crowbar device is often an ***SCR***; once triggered it latches on. A typical use is across the output of a ***buck regulator***. When the crowbar is activated, it is necessary to have a power or current limiting device somewhere or there will be a big bang. Typically a fuse would be put in series with the MOSFET; the crowbar will blow the fuse, thereby protecting both the circuit and the crowbar.

CURIE POINT … is the temperature at which a ferromagnetic material loses most of its permeability (μ_r decreases sharply). There are soldering iron bits [tips] that have a pellet of ferromagnetic material at their base. These are used to hold a reed switch closed by means of a magnetic field. When the pellet gets sufficiently hot its μ_r drops rapidly and the reed switch opens, disconnecting the heater current. This gives a very inexpensive control system to keep the soldering iron tip at the correct temperature. The drawback is that there is a current surge of several amps as the switch opens and closes; this can interfere with nearby circuitry. You can actually hear your soldering iron click and see a spike appear on your scope as you are working!

CUTOFF FREQUENCY: A ***waveguide*** will only pass electromagnetic energy above a critical minimum frequency known as the *cutoff frequency*, also known as the *dominant mode*. For a rectangular air-filled waveguide, the cutoff wavelength is double the longest side. Since the dielectric is air, the cutoff frequency is determined from the length of the longest side of the guide using:

$f_C = \frac{c}{2L}$, where *c* is the speed of light, and *L* is the length of the long side of the guide. Thus a 10 cm wide waveguide will only (easily) pass frequencies higher than 1.5 GHz.

The term *waveguide beyond cutoff* is the situation in which wavelengths much larger than the cutoff wavelength are being used. This means that frequencies much lower than the cutoff frequency are being used.

In cylindrical waveguides {pipes} the dominant mode is $TE_{1,1}$.This has a wavelength of $\lambda_C = 1.706 \times \text{diameter}$, which is the cutoff wavelength. The cutoff frequency for a 10 cm diameter pipe is therefore 1.758 GHz. The attenuation beyond cutoff is: [22]

$$\boxed{\alpha = \frac{2\pi}{\lambda_C}\sqrt{1-\left(\frac{\lambda_C}{\lambda}\right)^2}\ \text{nepers / metre}}$$

[22] E.G. Linder, 'Attenuation of Electromagnetic Fields in Pipes Smaller Than the Critical Size', in *Proceedings of the Institute of Radio Engineers*, 30 (Dec 1942), pp. 554-556.

At frequencies 10× or more lower than the cutoff frequency, the error resulting from neglecting the square root term becomes less than 0.5%. Thus to a good approximation:

$$\boxed{\alpha = \frac{2\pi}{\lambda_C}\ \text{nepers / metre} = \frac{2\pi}{\lambda_C} \times 8.6859\quad \text{dB / m} = 32\,\text{dB / diameter}}$$

A circular hole through a conducting enclosure as deep as its diameter will give at least 32 dB attenuation. In practice ambient RF fields inside the enclosure will not couple into the hole very effectively so the attenuation figure given is conservative.

As a rough rule of thumb, the waveguide beyond cutoff attenuation is roughly 30 dB/(longest side), regardless of the shape of the opening.

dBc … is dB relative to the **c**arrier [the principal frequency]. When measuring ***SFDR***, for example, the readings would be taken relative to the amplitude of the fundamental.

dBFS … is dB relative to **F**ull **S**cale. Full scale is always the largest input signal that can be applied to an input without clipping.

dBi … is a unit for antenna power gain relative to a *lossless isotropic antenna*, a fictitious creation which would radiate power equally in all directions. Because it is power gain, use 10× the log term not 20×. Rather than using the idea of a lossless isotropic antenna, one could instead consider the spatial mean of the power radiated by the real antenna; this amounts to the same thing as using the lossless isotropic antenna.

dBm … is dB relative to 1 mW in the impedance level of the system (usually 50 Ω). 0 dBm is therefore 223.6 mV RMS in a 50 Ω system. Typically used in RF systems.

dBu … is dB relative to 1 mW in a 600 Ω system. The *u* comes from *unloaded*. Typically used in audio systems when measuring voltage levels.

DECIMATION … means “discarding a large part of”, particularly with regard to data points. Consider an ADC sampling at 1 giga-samples per second in a ***DSO***. [GS/s is the generally accepted form of giga-samples per second]. Suppose you get this data rate on the top timebase of a DSO. If you turn the timebase down one position, the effective sample rate may now be 500 MS/s. Turn the timebase down further and the effective sample rate will keep reducing.

There are two ways of reducing the sample rate; either the ADC can be made to sample at a lower speed, or the excess data can be thrown away. The action of throwing away the unwanted data points is known as *decimation*. Decimation of the data is preferable to slowing down the ADC clock, because the ADC is kept working in the same way regardless of the timebase. This keeps the offset, gain, and the operating temperature the same. Additionally, the extra data can be used to find peaks {maxima} and troughs {minima} in the input signal.

DECONVOLUTION … is the inverse process of convolution. In the time domain an input signal, I(t), can be combined with the system impulse response, H(t), to give the output response, O(t). This combination is represented by the equation: $O(t) = I(t) * H(t)$

The * symbol is a shorthand notation for the convolution integral:

$$\boxed{I(t) * H(t) \equiv \int_{-\infty}^{\infty} I(t-\tau)\cdot H(\tau)\cdot d\tau}$$

The * symbol typically means “multiply” in many computer languages, so consider, by context, whether the * is being used as an ordinary multiplication or as a convolution.

For pulse response situations it may be necessary to *deconvolve* the input signal from the output signal in order to establish the true response of the system. An example of the use of this would be to measure the pulse response of a system with and without an additional external filter. The response of the filter itself can then be determined by deconvolution. This is non-trivial unless the filter risetime is several times slower than the system response.

DIMENSIONAL ANALYSIS: Quantities of different types cannot be added; you cannot add volts to amps, or watts to ohms. You must have the same units added or subtracted from each other. Whilst you could decompose a formula into the fundamental units mass [M], length [L], time [T], and current [I] it is quicker and easier to keep a check of the electrical units of current, voltage, impedance &c. The square brackets mean *the dimensions of*.

Dimensional analysis is not something new. It was originated in 1822 by J.B. Fourier; [23] widely known for his development of the *Fourier Transform*.

DIPLEXER … is a three-port device used for combining two signals for passage down the same path in the same direction. An example would be a dual band-pass filter with a common feed to both filters, usually made from linear passive components. The idea is that the two band-pass filters have non-overlapping pass bands, with excellent isolation between the bands. Thus the key parts of the definition are two signals in the same direction sharing a common path, such as an antenna. The same device could be used for simultaneous transmit and receive using the same antenna at slightly different frequencies, but in this case it would usually be called a *duplexer*. The key diplexer specs are insertion loss in each of the two pass-bands, band-to-band isolation, and the return loss at each port.

DIPOLE ANTENNA: A resonant centre-fed half-wave dipole is not exactly $\lambda/2$ long. When it is $\lambda/2$ long (and with thin antenna rods) the impedance is $Z = 73 + \mathrm{j}42.5$. A 5% frequency correction for the *end-effect* is needed, making the dipole resonant when $L \approx 0.475\lambda$; the impedance is then a pure resistance of ≈67 Ω, reducing by a further few percent for a factor of ten increase in the antenna rod diameter. Notice how sharp the resonance is in terms of the reactance change for a 5% shift in frequency.

When a horizontal half-wave dipole is less than a quarter wavelength from the ground, or any conducting plane, the radiation resistance drops almost linearly with distance to zero ohms at zero height from the conducting plane. The half-wave dipole needs to be further than a half wavelength from the ground in order to make the variation of radiation resistance within a ±12 Ω band of the free space value.

A dipole needs to be driven differentially at the centre point. Since RF signal generators are ordinarily single-ended, it is usual to have a ***balun*** right up next to the dipole; the balun converts the single-ended input signal into a differential signal for the antenna.

A dipole transmits (or optimally receives) linearly polarised electromagnetic waves, the plane of polarisation being in the same plane as the rods (elements).

DIRECTIVITY (ANTENNA): An isotropic radiator is a convenient theoretical device, but cannot exist in practice (except like a sun, in free space). All real antennas put out more power in one or more directions compared to other directions. This idea is quantified by the term *directivity*.

$directivity = \dfrac{\text{peak radiated intensity}}{\text{mean radiated intensity}}$, where it is understood that the terms peak and mean are related to the spatial distribution of the intensity pattern, not the time related values at a fixed position. The term *radiated intensity* is used as a power per unit solid angle and as such it is not affected by the distance from the antenna. If it is stipulated that the measurements are done at the same distance, then the directivity can be written as:

$$\boxed{directivity = \frac{\text{spatial peak radiated power flux}}{\text{spatial mean radiated power flux}}}$$

Notice that the directivity is not affected by the efficiency of the antenna or the mismatch at its input. The spatial mean power flux is equivalent to the idea that the total radiated power is coming from a

[23] J.C. Maxwell, 'On the Measurement of Quantities', in *A Treatise on Electricity and Magnetism*, 3rd edn (Clarendon Press, 1891; repr. Dover, 1954), pp. 1-2, Vol I.

lossless isotropic radiator. The directivity is therefore always higher than the antenna gain because of the *radiation efficiency* η. (See ***radiation resistance***.)

$$\boxed{\text{directivity} = \frac{\text{antenna power gain relative to a lossless isotropic radiator}}{\text{radiation efficiency}}} \qquad D = \frac{G_i}{\eta}$$

An isotropic radiator therefore has a directivity of 1, also stated as 0 dB isotropic, 0 dBi.

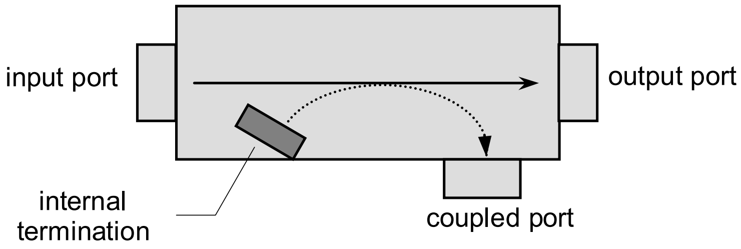
Loops and dipoles that are small compared to $\lambda/4$ have a directivity of 1.5 (1.76 dBi).

A $\lambda/2$ dipole has a directivity of 1.64 (2.15 dBi).

A 6 element (linear) *Yagi-Uda* antenna has a directivity of about 16 (12 dBi).

Highly directional antennas can have directivities of not only thousands, but millions.

DIRECTIONAL COUPLER … is a 3-port device typically inserted into a waveguide or transmission line system in order to be able to measure the amplitude of either the forward travelling (incident) wave or the reverse travelling (reflected) wave.



The signal is connected through the main I/O ports. This path has a defined attenuation of between 0.5 dB & 2 dB. The coupled port is often connected to a Z_0-matched RF voltage or power measuring device.

In the above diagram, the coupled port gives an attenuated version of the left-to-right ('input' to 'output') travelling wave.

There are six key specs:

1) Main-line insertion loss
2) Coupled port attenuation of forward travelling signal
3) Directivity
4) Operating frequency range
5) Coupling flatness
6) Mismatch (VSWR) at the I/O ports

Ideal Insertion Loss	Ideal Coupling Factor
0 dB	∞ dB
0.1 dB	> 16.4 dB
0.5 dB	> 9.6 dB
1.0 dB	> 6.9 dB
1.5 dB	> 5.3 dB
2.0 dB	> 4.3 dB
3.0 dB	> 3.0 dB

One octave[†] of frequency span is common for inexpensive couplers. Directional couplers are passive devices, so there is no power gain. Hence a low main-line insertion loss means the coupled port attenuation has to be high. For an ideal lossless coupler, if the main-line power gain is *M* and the coupled line power gain is *C* then $M + C = 1$. (This equation is not using decibels).

Thus $C = 1 - M$. In decibel form this equation becomes $C_{dB} = -10 \cdot \log_{10}\left(1 - 10^{\frac{-M_{dB}}{10}}\right)$

Directivity is the amplitude reduction at the coupled port when signal power is applied first at the 'input port' and then at the 'output port', with the other I/O port terminated correctly. <20 dB is poor; 30 dB is average; 40 dB is good; ≥45 dB is excellent. With a scalar measurement system the directivity is a serious limitation to the measurement uncertainty. With a vector measurement system, the directivity error can be calibrated out.

[†] A factor of two such as from 100 MHz to 200 MHz

If the internally terminated port is brought out to the user, the resulting device can be called either a 4-port directional coupler or a dual-directional coupler. Both coupled ports have to be well Z_0-matched or the directivity will suffer. In any case the best directivity achieved with a dual directional coupler will be at least 6 dB worse than an equivalent single version.

A directional coupler can also be used as a signal combiner. The two sources are connected to the ‘coupled port’ and the ‘output port’, with the combined signal coming out of the ‘input port’.

DISPERSION … causes increased softening of the fast edges of a rectangular pulse as the pulse travels down a cable or waveguide. If the component frequencies of a pulse travel at different group velocities in a particular medium, they will tend to separate-out as the distance travelled in the medium is increased. The medium could be a coaxial cable for electrical signals, an optical fibre for light, or a waveguide for microwaves and mm-waves. Hollow waveguide is highly dispersive. In an optical fibre the change of velocity with wavelength (colour) would be called chromatic aberration.

Modal dispersion occurs where the medium can support multiple transport modes. In a multi-mode optical fibre, for example, the limiting distance of operation is not due to attenuation. At 850 nm the attenuation in an optical fibre would be ≈3 dB/km, whereas at 1300 nm this drops to <0.5 dB/km. However the modal dispersion causes one symbol to blur into another, giving *inter-symbol interference*. Clearly the two factors involved are distance and speed of operation, the two being multiplied together to give a figure of merit for multi-mode fibre. A typical value is 200 MHz×km, where “MHz” really means mega-bits per second. Dispersion in single-mode fibres has a different mechanism, making the limit in terms of GHz ^{2}×km, a typical value being 5000 GHz ^{2}×km. This gives 800 km at 2.5 Gb/s, but only 50 km at 10 Gb/s.

DISTORTION … means that the system output is unintentionally of a different shape or nature to that which could ideally be expected from the given input. It is difficult discuss a general system without making the description almost incomprehensible. I will therefore use an amplifier as an example.

The reason for the part about the “shape or nature” being the same is that an amplifier will have gain, so the output signal will be bigger than the input signal. For attenuators and transmission lines, the output signal will be less than the input signal. (Except for special case of high-Z load. See p 561).

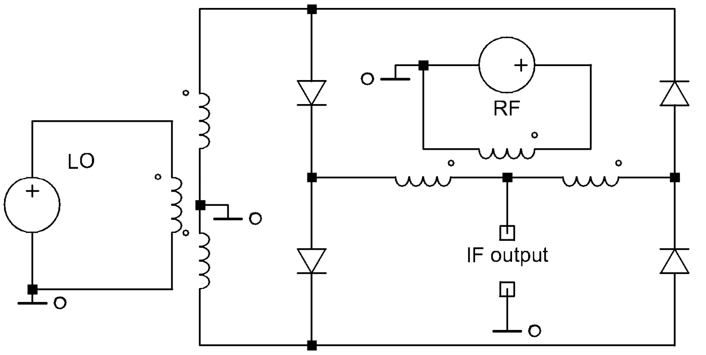
If you measure the frequency response of an amplifier, the gain will not be constant with frequency. This *flatness error* of the amplifier creates distortion of any non-sinusoidal signal, but if the response is measured by a swept frequency input, the resulting flatness error would not ordinarily be referred to as distortion.

The same amplifier could be driven from a rectangular waveform and the result viewed on a scope. The (distorted) output waveform could then be characterised by such measures as *rise-time*, *overshoot*, *pre-shoot*, *ringing*, *droop* &c. You would not characterise a time-domain result by an otherwise unqualified number such as “3% distortion”; you would need to be more specific, such as 3% overshoot in the first 10 ns.

Non-linear distortion could be quantified as *harmonic distortion*, two-tone ***intermodulation distortion***, or change of waveshape with input signal amplitude. You would pick whichever test method gave the most useful results according to your application. Harmonic distortion in an audio system is obviously important, but if the system has a 20 kHz bandwidth, the third harmonic distortion of a 15 kHz signal would seem to be unimportant. In practice the distortion is important, but is not shown up by the harmonic distortion test. A two-tone intermodulation test, using say 15 kHz and 16 kHz signals, would then be useful; the 1 kHz difference frequency being nicely within the system pass band.

DOMAIN: For electronics work, the most appropriate English definition of *domain* is the figurative one, meaning a place of thought or action. Typically use is made of the terms *time domain* and *frequency domain*. Think of this as plotting the measured or observed data points on a graph. It is the horizontal axis, the *independent variable*, which gives the domain. A scope measures against time and is therefore a time domain instrument. A spectrum analyser is then clearly a frequency domain instrument. A *Bode plot* is in the frequency domain. Pulse response is a time domain measure.

DOUBLE-BALANCED MIXER (DBM): A local oscillator, LO, is mixed with a smaller RF input to produce sum and difference frequencies at the Intermediate Frequency, IF, output.



The mixer is usually designed for a specific small range (6 dB) of local oscillator signal amplitudes and a wide range of input frequencies (decades). The signal level is optimised to alternately turn on first one string of diodes then the other.

Since the transformer secondaries are well matched, and the diodes are also well matched, the mid-points of the diode chains are effectively shorted to ground on alternate half-cycles of the local oscillator. The diodes act as low impedances switches to the signal ground, switching the polarity of the RF signal fed to the output.

Optimum LO signals levels can be selected in the range from 7 dBm to 23 dBm, although any particular mixer may work reasonably well if the LO level is up to 5 dB below its stated optimum level. The RF input level must always be kept 6 dB lower (half the voltage) than the LO level. The resulting IF output level is lower than the RF input level by about 6 dB, the exact figure being called the *conversion loss*.

The degree of balance within the device affects the amount of local oscillator feedthru to the IF output. This will vary according to the mixer quality from 20 dB isolation (poor) to 50 dB (very good). Notice that the IF output is DC coupled and will therefore work all the way down to DC. The LO and RF inputs, on the other hand, are transformer coupled and therefore only work over some restricted range of input frequencies.

DROP-OUT VOLTAGE … is the voltage at which a circuit or a component stops working to spec.

One specific definition is the voltage difference required from input to output on a linear voltage regulator for it to work correctly as a regulator, also referred to as the *headroom*. The headroom required increases with current drawn from the regulator. Older regulator designs need perhaps 1 V to 3 V input-output voltage to operate correctly. Modern LDOs {Low Drop-Out regulators} can function correctly with as little as a few tens to a few hundreds of millivolts of headroom.

Another specific definition is for the *turn-off* voltage of a relay, also known as *must-release* voltage.

DRY CIRCUIT: An electro-mechanical switch, a relay for example, or indeed a mechanical switch, suffers from contact resistance problems when the load being switched is low in terms of voltage (<100 mV) *and* in terms of current (<10 μA). These figures can only be approximate as they vary according to the contact materials, the contact pressure and the amount of contaminants present. Under these low-load conditions you have a *dry circuit*. Higher voltages or currents break through the contamination layer and make a lower resistance contact; the current is said to "wet" the contact. It is therefore necessary to specify both the maximum and minimum loads that a relay or switch can reliably deal with.

Reed-relays can switch arbitrarily low voltages or currents because they are ***hermetically sealed***. Many relay datasheets do not specify a minimum voltage rating for the contacts. This could be because the manufacturer is lazy, ignorant, or does not intend the relay to be used for small-signal applications. If used on such a low level signal, the relay may perform badly.

Mercury-wetted reed switches have (liquid) Mercury inside them and are differentiated from "dry reed switches" that have no Mercury. Because of their excellent bounce characteristics, mercury-wetted reed switches can be used as primary standards of flat pulse waveforms for repetition rates below 40 Hz.

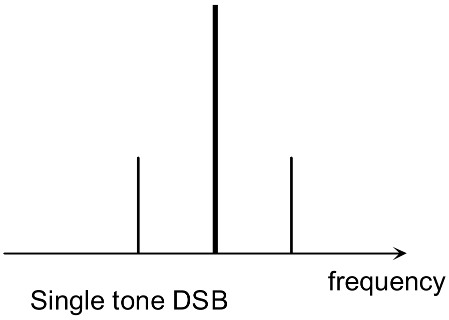
DMMs are available to measure resistance under 'dry' conditions. One manufacturer specifies a maximum open-circuit source voltage of 20 mV on the resistance measurement so that any contamination film is not punctured.[24]

DRY JOINT: The solder does not "wet" the conductor surfaces and the joint is faulty, both electrically and mechanically. This is caused by dirty conductor surfaces and/or insufficient/inappropriate flux. The flux must eat through any oxide or contamination film on the conducting surfaces and must therefore be matched to the conductors. The flux required for soldering steel is different to that used for soldering tin.

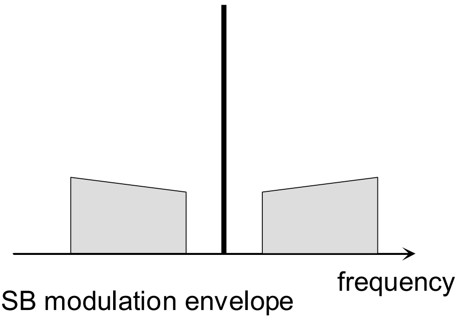
If the conductors are large and the heat source (usually a soldering iron) is too small, the solder does not get hot enough to flow around the joint; this gives a *cold soldered* joint. Again the joint is neither mechanically nor electrically sound {correct; reliable}.

DSB: **D**ouble **S**ide-**B**and

On a spectrum analyser display DSB appears as tones (or modulation envelopes) mirrored about the carrier. The horizontal scale is frequency, and the vertical scale is dB amplitude, both as you would get on a spectrum analyser display. DSB signals are naturally produced by amplitude modulation and can be detected (demodulated) using a simple diode/capacitor filter to give the "envelope" of the waveform.



Single tone DSB



DSB modulation envelope

DUPLEX OPERATION: Simultaneous transmission in opposite directions is known as duplex operation, or sometimes *full-duplex*. *Half-duplex* then means alternate send and receive down the same path. A *duplexer* and a *diplexer* could physically be identical components, the intended purpose making the distinction in the name. If the signals are travelling in the same direction simultaneously then that would be *diplex*, whereas opposite directions would be *duplex*.

DUTY CYCLE … is the proportion of a cycle for which the power is "on". This is particularly useful for rectangular waveforms, the duty cycle times the peak power giving the mean power: $P_{MEAN} = D \times \hat{P}$. For this reason, duty cycle is more commonly used than *mark/space ratio*. Because power is voltage squared over resistance, the RMS voltage of a rectangular waveform is related to the peak voltage by the square root of *D*. $\boxed{V_{RMS} = \hat{V} \cdot \sqrt{D}}$

The definition of duty cycle can be extended for communication systems where there may be many rectangular pulses in some sort of pattern, repeating over a particular frame period. The duty cycle is still the total on-time divided by the cycle time and still gives the mean power as above.

There is a nasty variant definition related to equipment which is not rated for continuous operation.[25] For example a hand-held power drill might have a rating limit such as 20 minutes total use in any 2 hour period. Avoid using this definition.

DVM Scale Sizes: A 3 digit meter will display up to a count of ±999 with a decimal point placed before/after any digit. A 3½ digit meter will give a full scale count of ±1199, ±1200, ±1999 or ±2000.

[24] Keithley, '3.3.5: Dry Circuit Testing', in *Low Level Measurements*, 5th edn (USA: Keithley Instruments Inc, 1998), pp. 3.23.

[25] 'Duty Cycle' in Federal Standard 1037C (1996). Telecommunications: Glossary of Telecommunications Terms.

The common feature is that the first digit doesn't have the full extent of the others; usually it can only be either zero (blanked) or one.

DYNAMIC RANGE … is that discernable {measurable} range of input signal over which the response of a device is specified. This is generally taken to be from the noise level up to some agreed level of non-linearity. For an RF amplifier, this upper limit might be the *1 dB compression point*, by which is meant the output signal level at which the gain is 1 dB lower than small-signal gain.

For an ADC, the dynamic range would be from 1 LSB up to the maximum possible conversion range. An 8-bit ADC would have a maximum possible dynamic range of $20 \cdot \log_{10}\left(\frac{2^8}{1}\right) = 48\,\mathrm{dB}$

A slight warning is required here. The upper and lower limits of the dynamic range should, in principle, be measurable at the same time. On an ADC this might be seen by looking at a 1 LSB signal riding on a near full-scale DC signal. On a spectrum analyser, both signals could be injected at the same time and you would genuinely be able to measure them simultaneously. If it were not for this requirement then low resolution systems with auto-ranging inputs would appear to have very high dynamic ranges. Take the example of an auto-ranging 3-digit DVM with ranges from 10 mV to 1000 V. Let's suppose the noise is 0.03 mV on the 10 mV range. It would not be correct to say that this meter has a dynamic range of $20 \cdot \log_{10}\left(\frac{999}{0.03 \times 10^{-3}}\right) = 150\,\mathrm{dB}$

In reality the best dynamic range would be achieved on the top range, where the relative amount of noise would be the lowest. Hence the dynamic range would actually be $20 \cdot \log_{10}\left(\frac{999}{1}\right) = 60\,\mathrm{dB}$

There is a complication with log-compression amplifiers. It is no longer very meaningful to consider the simultaneous application of two signals and the response is never "linear". In situations like this, it is not uncommon for manufacturers to 'invent' their own definitions for terms.

EARLY EFFECT … is the effect of *base-width modulation* in bipolar transistors, caused by changing the collector-emitter voltage.[26] Specifically, the collector-base junction gets wider with increased reverse bias and this makes the base effectively thinner. The result is that increasing the collector voltage, for a given base-emitter bias, increases the collector currently slightly. This is important because it means that a transistor wired as a current source has a finite output resistance which reduces with increasing collector current.

The model for this effect includes a voltage generator, the voltage being called the *Early Voltage*, V_A. The large signal equation is: $I_C = I_S\left(\exp\left[\frac{V_{BE}}{V_T}\right] - 1\right)\left(1 + \frac{V_{CE}}{V_A}\right)$

V_A can easily range from 10 V to 150 V for different types of transistor so you must make sure that any SPICE simulation work you do includes a valid value for VAF, the *forward Early voltage*. Without this, the finite output impedance of current sources will not be modelled and your actual circuit will not perform as well as the simulation.

Doing a partial differentiation of the large signal equation with respect to V_{CE} gives:

$$\frac{\partial I_C}{\partial V_{CE}} = I_S\left(\exp\left[\frac{V_{BE}}{V_T}\right] - 1\right)\left(\frac{1}{V_A}\right) = \frac{I_C}{\left(1 + \frac{V_{CE}}{V_A}\right)}\left(\frac{1}{V_A}\right) = \frac{I_C}{V_A + V_{CE}}$$

[26] J.M. Early, 'Effects of Space-Charge Layer Widening in Junction Transistors', in *Proceedings of the Institute of Radio Engineers* (Nov 1952), pp. 1401-1406.

The small-signal output resistance is therefore given by:

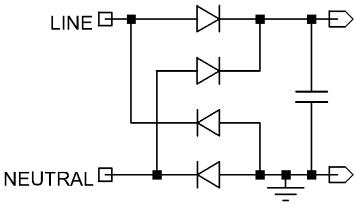
$$\boxed{r_o = \frac{\partial V_{CE}}{\partial I_C} = \frac{V_A + V_{CE}}{I_C}}$$

For a typical case of an Early voltage of 95 V, a collector-emitter voltage of 5 V, running at 10 mA collector current, the output resistance is only 10 kΩ. The output resistance is increased tenfold if a resistor is put in series with the emitter such that around a 0.5V drop is created due to the current.

Remember that these output resistances may only be achieved at DC and up to a few hundreds of kilohertz. When the current gain of the transistor starts to roll-off, the output resistance also rolls-off at the same rate. This roll-off frequency is easily seen to be the transistor's current gain-bandwidth product divided by its low-frequency gain. This is the β-cutoff frequency of the transistor: $f_\beta = \frac{f_T}{\beta}$

EARTH: The planet you are on is called Earth. The soil is also referred to as earth. The soil is relatively conductive (resistivity between 100 Ω·m and 10 kΩ·m) and main power distribution grids are referenced to earth via buried conductors. Domestic supplies may have an earth feed from the power supplier in the form of a connection to a steel wire armoured (SWA) cable or the newer aluminium wire armoured (AWA) cable. In this case the correct name is the *protective conductor*. Alternatively there may be a pipe or rod driven into the ground as an *earth electrode*.

Circuit diagrams and circuit simulators often use the AC mains earth symbol when really they mean a signal 0 V connection; this may or may not be earthed. You have to be aware of this and understand which one is being talked about. If you use the earth symbol on a circuit with mains related circuitry on it, then make sure you mean earth and not signal 0V, or there will be a big bang!



This is an actual example of part of a circuit in an application note for an off-line switched-mode power supply from a major semiconductor supplier. An earth symbol has been used instead of a chassis symbol or a 0 V symbol.

If you wire this circuit up as drawn, it will blow up … *violently*!

In the USA the term *ground* is used in place of earth.

EIRP: **E**ffective (**E**quivalent) **I**sotropic(ally) **R**adiated **P**ower. EIRP can be used to compare peak signal strengths from different transmitters.

$$EIRP = \text{accepted power} \times \text{efficiency} \times \text{antenna gain relative to a lossless isotropic antenna}$$

A more directive transmitting antenna gives a higher received field strength, making the transmitter appear more powerful. EIRP is measured in watts.

$$\boxed{EIRP = \text{radiated power} \times \text{directivity}}$$

$$\boxed{EIRP = \left(\frac{4\pi d}{\lambda}\right)^2 \cdot \frac{P_r}{G_{Pi}}}$$

To measure the EIRP of a transmitter, use the peak received power (P_r), the *practical gain* (G_{Pi}) of the receiving antenna (not in dB), the operating wavelength (λ), and the distance (*d*) from the transmitting antenna.

ELECTRET … is the electrostatic equivalent of a permanent magnet. It is possible to make an electret by melting an insulator then allowing it to solidify in a strong electric field, say >100,000 V/m. This does not necessarily make a good or permanent electret, but it does illustrate the principle of having an electric polarisation built into the material. The name was coined by Heaviside in 1885.[27] The first good, permanent electrets were made around 1922.

Electret microphones are inexpensive solid-state devices which require a small DC bias and a pre-amplifier when replacing moving coil microphones. They are ideal suited to harsh environments due to their homogeneous construction.

[27] O. Heaviside, 'Electrization and Electrification. Natural Electrets.', in *The Electrician 1885* (repr. Electrical Papers vol I) (1892, AMS 2001), pp. 488-493.

ELLIPTIC FILTER … is a filter type, named because its poles lie on an ellipse in the complex plane. Also called a ***Cauer-parameter*** filter. By way of comparison, the poles of a Butterworth filter lie on a circle in the complex plane.

ELLIPTIC INTEGRALS … are a group of analytically insoluble integrals which appear in electric and magnetic field problems. The name *elliptic integral* comes from their original purpose, evaluating arc-length of ellipses (Legendre, 1811). Typically complex field patterns are solved using *conformal mapping*, giving results in terms of elliptic integrals. One then needs either tables of values [28] or approximation formulae. Calculation of inductances, capacitances and characteristic impedances often turn out to involve elliptic integrals.

$$\boldsymbol{K}(k) \equiv \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} \equiv \int_0^{\frac{\pi}{2}} \frac{d\theta}{\sqrt{1-\ k^2\sin^2(\theta)}}$$

Complete elliptic integral of the 1st kind.

$$\boldsymbol{E}(k) \equiv \int_0^1 \sqrt{\frac{1\text{-}k^2t^2}{1-t^2}} \cdot dt \equiv \int_0^{\frac{\pi}{2}} \sqrt{1-k^2\sin^2(\theta)} \cdot d\theta$$

Complete elliptic integral of the 2nd kind.

The standard forms using the *sine* functions are known as *Legendre's complete elliptic integrals*. Shorthand notation omits the parameter *k*, thus $\boldsymbol{K}(k) \equiv \boldsymbol{K}$ and $\boldsymbol{E}(k) \equiv \boldsymbol{E}$.

The more general form of these elliptic integrals has the upper limit as a parameter.

$$\boldsymbol{F}(\phi,k) \equiv \boldsymbol{F}(\phi) \equiv \int_0^{\sin(\phi)} \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} \equiv \int_0^{\phi} \frac{d\theta}{\sqrt{1-\ k^2\sin^2(\theta)}}$$

$$\boldsymbol{E}(\phi,k) \equiv \boldsymbol{E}(\phi) \equiv \int_0^{\sin(\phi)} \sqrt{\frac{1\text{-}k^2t^2}{1-t^2}} \cdot dt \equiv \int_0^{\phi} \sqrt{1-k^2\sin^2(\theta)} \cdot d\theta$$

WARNING: Some tables and maths software give elliptic integrals in terms of the parameter *m*, where $m = k^2$.

The inverse functions of these elliptic integrals are *Jacobian elliptic functions*. Mathematicians consider the elliptic functions to be the natural forms, with elliptic integrals being the inverse functions.

Let $u \equiv \int_0^{\phi} \frac{d\theta}{\sqrt{1-\ k^2\sin^2(\theta)}}$, then *am*(*u*,*k*) is defined as that value of ϕ which makes the integral equal to the required value of *u*, *am* being short for amplitude.

Similarly, if $u \equiv \int_0^{x} \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}$ then *sinam*(*u*,*k*) is defined as that value of *x* which makes the integral equal to the required value of *u*, *sinam* being the 'sine of the amplitude' and usually being written in the abbreviated form *sn*(*u*,*k*).

In all cases of elliptic integrals and elliptic functions, when the parameter *k* is omitted, its existence is just 'understood'.

EMF: Electro **M**otive **F**orce. Used to refer to an open-circuit (un-loaded) voltage, rather than the measured terminal voltage. EMF is measured in volts and is therefore is not a mechanical force, although the associated electric field does exert force on charged particles.

Since EMF is not a force, the term *electromotance* has been used in its place in some older text

[28] M. Abramowitz, and I. Stegun, 'Chapter 17, Elliptic Integrals', in *Handbook of Mathematical Functions* (National Bureau of Standards, 1964; repr. Dover Publications, 1965).

books. However, the term electromotance never really caught on, so the recommendation is to use EMF as the letter abbreviation, but don't express it in its unabbreviated form.

In the world at large, the term EMF can stand for "Electric & Magnetic Field" and also "Electromagnetic Field". Avoid using these last two definitions.

EMISSIVITY: Radiated heat flow from a body is defined by the *Stefan-Boltzmann Law*:

$$\boxed{P = \varepsilon \sigma T^4}$$

where *P* is the heat flow per unit area [W/m²]

T is the absolute temperature [°K]

σ is the Stefan-Boltzmann constant [56.7 W/(m²·°K4)]

ε is the emissivity; a dimensionless value, $0 < \varepsilon \leq 1$.

A *Black Body*, an ideal radiating or absorbing surface, would have an emissivity of 1. In the real world, ε is a function of the wavelength of the radiation being measured, the temperature and the direction of the measurement relative to the surface of the body under consideration. Emissivity is equal to absorptivity at any given wavelength.

This table of emissivities gives an idea of the range of possible variation.[29] It demonstrates the importance of calibrating any infra-red sensing device to the material of the surface being viewed.

Material	Emissivity
Polished silver	0.025
Polished copper	0.05
Polished brass	0.06
Polished chromium	0.075
Aluminium sheet	0.10
Zinc (as galvanised iron sheet)	0.23
Concrete tiles	0.63
Rough lime plaster	0.91
Rough red brick	0.93
Rough steel plate	0.96
Black lacquer	0.98

Although the highest value of emissivity for a simple surface is around 0.98, the total radiant power leaving a surface can be considerably higher than σT^4. The total radiant power, the *radiosity*, is equal to the sum of the radiant power given by the Stefan-Boltzmann Law and the reflected power; the reflected power being the incident power multiplied by the reflectivity of the surface.

α, the absorptivity, is the per-unit amount of incident power absorbed by a surface. ρ, the reflectivity, is the per-unit amount of incident power reflected from a surface. τ, the transmissivity, is the per-unit amount of incident power passing through a surface.

$$\boxed{\alpha + \rho + \tau = 1}$$

It is therefore possible for a shiny surface to appear much hotter than it really is, if measured with an infra-red sensor.

ERROR FUNCTION: Abbreviated as erf(x).

$$erf(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-y^2\right) \cdot dy = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n \cdot x^{2n+1}}{n!(2n+1)}$$

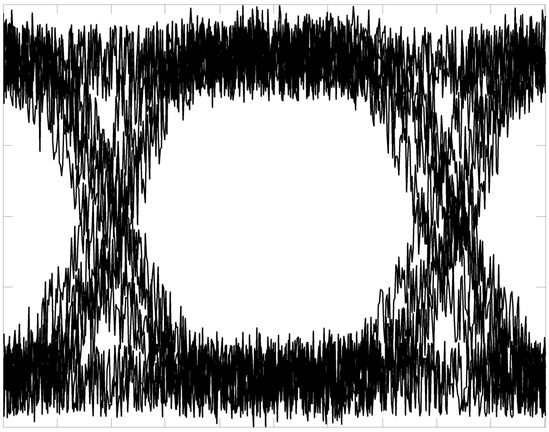
$erf(x) = 2 \times cnorm\left(x\sqrt{2}\right) - 1$, where $cnorm(x)$ is the cumulative probability function of the Gaussian distribution with mean 0 and variance 1. The complementary error function, erfc(x), is:

$erfc(x) = 1 - erf(x)$ $erf(0) = 0$ $erf(\infty) = 1$

The erf(x) and erfc(x) functions have also historically been defined without the $\frac{2}{\sqrt{\pi}}$ factor.

[29] J.R. Simonson, *Engineering Heat Transfer*, 1st edn (Macmillan Press, 1975; repr., 1981).

EYE DIAGRAM: If you trigger a scope on the clock of a data stream and use trace persistence {all previous trace positions are retained on the screen}, you get rising and falling edges corresponding to the changes in the data steam. Because of the finite edge speed, clock-to-data jitter, noise, dispersion and *inter-symbol interference*, you get an "eye" in the middle of the display where there is no trace. Ideally the eye will be "open" demonstrating a relatively noise free system. If there is a lot of system noise, the eye will be closed up and there will be a higher probability of an error in the received data. Thus the largest vertical opening of the eye, as a function of the overall trace height at that point, is a semi-quantitative measure of the system Bit Error Ratio (***BER***).



If you look hard at this picture you can just see a noisy rising edge and a noisy falling edge overlaid. It is important that the eye diagram is produced from a Pseudo Random Binary Sequence (PRBS) rather than simply repeating a 1010 pattern. The different high and low periods in the PRBS will reduce the opening of the eye due to time-constant effects such as *droop*. It also is important that adjacent channels be exercised at the same time, but with uncorrelated patterns, allowing crosstalk effects to be taken into account. These data related effects are grouped together under the name *inter-symbol interference*.

FARADAY's LAW OF INDUCTION ... is often expressed as $V = -N \cdot \frac{d\phi}{dt}$. The induced voltage, *V*, is equal to the time rate of change of the flux, ϕ, passing through *N*-turns of a fixed electrical circuit. This law was discovered by Faraday in 1831.

Faraday's law is sometimes stated as being a voltage generated by the time rate of change of *flux linkages*. Rate of change of flux linkage can easily be misinterpreted, however. If you take:

$$V = -\frac{d(N\phi)}{dt} = -N \cdot \frac{d\phi}{dt} - \phi \cdot \frac{dN}{dt}$$

Consider a transformer with a variable tap point, as you find, for example, on power auto-transformers. If a steady "DC" flux is passing through the core, you do not get voltage appearing at the wiper as you move it. Thus the term $\phi \cdot \frac{dN}{dt}$ does not have a direct physical existence. However, if the core is energised with an AC flux, moving the wiper will change the output voltage.

Numerous (un-workable) patents have been applied for which rely on misinterpretations of Faraday's law. The faulty reasoning is explored nicely in a monograph by Bewley.[30]

FEEDBACK ... is the process of changing the gain of a system by taking some portion of the output signal and allowing it to return to the input.[31] If this feedback increases the overall gain it is called *positive feedback*.

Positive feedback can be used to increase gain; decrease bandwidth, and therefore increase selectivity; produce a ***Schmitt trigger***; produce oscillations.

Negative feedback is used to stabilise gain against component variation, increase bandwidth, reduce overshoot & ringing, and reduce distortion.

FFT: **F**ast **F**ourier **T**ransform. This is a computationally efficient technique which takes a time domain waveform (a set of data points, sampled at a regular interval) and converts it to a frequency domain waveform (a set of resultant frequency points). The original data set has to be an integer

[30] L.V. Bewley, *Flux Linkages and Electromagnetic Induction* (Macmillan, 1952).
[31] H.S. Black, 'Stabilized Feed-Back Amplifiers', in *Electrical Engineering*, 53 (Jan 1934), pp. 114-120.

power of 2. A Discrete Fourier Transform (DFT) is computationally slower, but produces the same sort of response with an arbitrary length of input data.

DFT requires of the order of N^2 operations

FFT requires of the order of $N \cdot \log_2(N)$ operations.

Thus an FFT gets more and more efficient, compared to a DFT, as the number of sample points, *N*, is increased. The FFT algorithm, rediscovered by Cooley and Tukey,[32] is a relatively recent addition to an old subject.

Time domain data generated by an ADC has a resolution of one LSB. Thus an *N*-bit converter has a resolution of one part in 2^N. From this result you might expect that an FFT of the ADC data would show a noise floor of:

$$-20 \cdot \log_{10}\left(2^N\right) = -20 \cdot N \cdot \log_{10}(2) \approx -6.02 \cdot N \quad \text{dBFS}$$

What actually happens is that the FFT spreads the noise out over the number of frequency bins available, meaning that a longer FFT gives a lower noise floor. The actual formula for the noise floor is therefore:

$$\boxed{\text{Noise Floor} = -\left[6 \cdot N + 10 \cdot \log_{10}(M)\right] \text{ dBFS}}$$

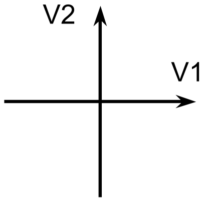
where *M* is the length of the FFT. It is more realistic to put *N* as the *effective number of bits* of the ADC. If an ADC is specified as *N* bits, you are doing very well (and operating at a fairly low signal frequency) if you manage to get *N*–½ effective bits from it, < N–1 effective bits being more usual.

FORM FACTOR … is defined as the ratio of the RMS value of a waveform to its *mean absolute* value. Another definition of Form Factor gives it as the RMS value divided by the 'full-wave-rectified mean'. This definition is better than using 'half-cycle mean', as it allows for waveforms with even-harmonic distortion. If *T* is the period of a cyclic waveform, the definition is unambiguously represented in mathematical form:

$$\boxed{Form\ Factor = \frac{\sqrt{\dfrac{1}{T}\displaystyle\int_0^T v^2(t) \cdot dt}}{\dfrac{1}{T}\displaystyle\int_0^T \left| v(t) \right| \cdot dt} = \frac{\text{cycle-RMS value}}{\text{cycle-mean-absolute value}}}$$

"If one dared to disregard proprieties of language, the form factor might otherwise be called the 'coefficient of peakiness' of the curve." [33]

FOUR-QUADRANT OPERATION: The definition of quadrant which fits best is "one of the four parts of a plane that is divided by two lines crossing at right angles".



Four-quadrant operation is when any combination of positive and negative signal is acceptable on the two inputs V1 and V2. If, for example, V1 can be either positive or negative, but V2 must remain positive, then that is *two quadrant operation*. If you have a two-quadrant multiplier, but you want to multiply two AC signals, then it is possible to use a DC bias on the input which cannot take reversed polarity signals. The DC bias has to exceed the peak excursions of the AC signal in order that the multiplier always sees a fixed polarity signal.

[32] J.W. Cooley, and J.W Tukey, 'An Algorithm for the Machine Calculation of Complex Fourier Series', in *Mathematics of Computation*, 19, no. 90 (Apr. 1965), pp. 297-301.

[33] J.A. Fleming, 'The Form Factor of Alternating-Current Curves', in *The Electrician*, XXXVI (Jan 1896), p. 338.

FWHM: **F**ull **W**idth **H**alf **M**aximum. A definition of pulse width measurement. The width is defined as the time between the 50% points of the pulse, one on the rising edge and the other on the falling edge.

GAIN MARGIN: In control theory, and in amplifier feedback systems in general, the system becomes unstable if the ***loop-gain*** reaches unity with 180° phase shift. The *gain margin* is the amount that the loop-gain is lower than 1 when the loop phase shift becomes 180°. The gain margin is easily seen on a ***Bode plot***. Given a forward gain factor of $G(j\omega)$ and a feedback factor of $H(j\omega)$, the Bode plot gives the magnitude of the loop-gain in decibels, $20\times\log_{10}|G(j\omega)\times H(j\omega)|$ dB. The gain margin is simply read off of the Bode plot magnitude scale at the frequency where the phase shift first reaches 180° (with increasing frequency). The gain has to be less than 0 dB, and the number of dB below zero is the gain margin.

On a system where the loop phase shift approaches 180° on a shallow curve, the phase margin is far more important than the gain margin. In any case a phase margin of at least 45° is desirable, as is a gain margin of at least 10 dB. These values are only a starting point for a design. The important thing to know is that a peaking frequency response in a closed-loop system is due to inadequate gain margin and/or inadequate phase margin.

GALVANIC ISOLATION … means "no direct current path". A transformer provides *galvanic isolation*. Any physical path, including isolating transformers and opto-couplers, will have a finite insulation resistance, although these can exceed 10 TΩ. Galvanic Isolation should be thought of in terms of *no intentional resistive path*, with a leakage resistance > 1 GΩ.

GAUSSIAN DISTRIBUTION … also known as the *Normal distribution*. It has a probability density function given by $p(x)=\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{x^2}{2}\right)$.

See the table given in the Useful Data section of this book, and ***central limit theorem***.

GAUSSIAN FILTER: The magnitude of the transfer function, *T*, of a Gaussian filter follows an $\exp\left[-k\cdot f^2\right]$ law. If this is expressed in terms of the 3 dB bandwidth, *B*, then you have:

$|T|=A\cdot\exp\left[-0.3466\left(\frac{f}{B}\right)^2\right]$, where *A* is the low frequency gain of the filter.

Note that the constant is $\ln\left(\sqrt{2}\right)=0.3466$. The step response of a Gaussian filter is:

$$\boxed{v(t)=V\times\left(0.5+0.5\times erf\left[\frac{t-t_0}{T_R}\times 1.81239\right]\right)}$$

Where V is the step amplitude, and $erf(x)$ is the ***error function***. At time t_0 the pulse is at the 50% point of the step. T_R is the usual 10% to 90% risetime of the pulse. This pulse edge is rotationally symmetric about its mid-point (50% point), unlike the ***Gaussian pulse***. The pulse response has been described as an *integrated Gaussian* because:

$$v(t)=V\times\left(0.5+\frac{1}{\sqrt{\pi}}\int_0^t\exp\left[-\left(\frac{t-t_0}{T_R}\times 1.81239\right)^2\right]\cdot dt\right)$$

An integrated Gaussian pulse differs from a Gaussian pulse of the same risetime and amplitude by ±4%FS. If the error function is not readily available for simulation purposes, then this next function will give a reasonable approximation to the correct response:

$v(t) = 1 - \exp\left(-0.362423\left(\frac{t}{T_R}\right)^{3.48}\right)$. This fits the integrated Gaussian pulse to with ±0.5%FS.

GAUSSIAN PULSE: is defined by:

$$v(t) = V \times \left(1 - \exp\left[-\left(\tfrac{t}{A}\right)^2\right]\right)$$

Using the usual 10% to 90% risetime, the unit Gaussian pulse is:

$$\boxed{v(t) = 1 - \exp\left(-1.42286\left(\frac{t}{T_R}\right)^2\right)}$$

The Gaussian pulse is not symmetrical about the 50% point. However, an *integrated Gaussian* pulse, the response of a ***Gaussian filter***, is symmetrical about this mid-point of the step.

GIBBS' PHENOMENON: The Fourier series reconstruction of a waveform does not converge when there is a discontinuity in the waveform. Thus summing the components of a square wave always results in both overshoot and preshoot, regardless of the number of terms used.[34] The amount of the overshoot is approximately 8.9%. This was publicised by Gibbs in 1899.[35] Increasing the number of terms in the Fourier series reduces the area of the overshoot by reducing its width. In other words, the overshoot has components at a higher frequency. Thus increasing the number of harmonics can increase the bandwidth of the signal to such a degree that the system response will filter out the overshoot. This is how an increased number of harmonics in a rectangular wave converges to give an arbitrarily good pulse response.

GLITCH … is a narrow pulse or transient spike on a waveform. A glitch can result from badly designed logic circuitry. A glitch can also be the result of a fast edge in one wire capacitively coupling onto another (*cross-talk*). A glitch is ordinarily considered to be too narrow to be valid. For example, a glitch on a common clock line may cause some gates to be clocked and others not to be clocked. It is so narrow that not all the gates can 'see' it. If the pulse is narrow, then it may also be too low in amplitude due to the finite slew rate of the signal. This is still a glitch, but it is more properly referred to as a ***runt pulse***.

GROUP DELAY: If a network is not going to distort a signal in any way, then it must attenuate or amplify signals of all frequencies by the same amount. This is a *necessary but not sufficient* condition. It is also necessary that each frequency should be delayed in time by the same amount as well, or a time domain signal, such as a pulse, will be dispersed.

$\boxed{Group\ \ Delay = -\frac{d\phi}{d\omega}}$ To get a constant group delay, the rate of change of phase with frequency has to be constant; this is the origin of the term *linear phase* filter.

GROUP VELOCITY: Electromagnetic energy flows at the group velocity, which is always smaller than the speed of light in that substance. CPW/coax are ideally non-dispersive. In hollow waveguide the group velocity is a strong function of the frequency, and hence wavelength, of the signal.

$$\boxed{Group\ \ Velocity = c \cdot \sqrt{1 - \left(\frac{f_C}{f}\right)^2}} \qquad \text{for } f > f_C$$

where *c* is the speed of light and f_C is the ***cut-off frequency*** of the (air filled) waveguide.

[34] L.O. Green, 'Fourier Synthesis: Defeating the Gibbs Phenomenon', in *Electronics World* (Highbury Business Communications), March 2003, pp. 48-51.

[35] E. Hewitt, and R.E. Hewitt, 'The Gibbs-Wilbraham Phenomenon: An Episode in Fourier Analysis', in *Archive for History of Exact Sciences*, 21 (1979), pp. pp 129-160.

HALL EFFECT: When Hall read Maxwell's "Treatise on Electricity & Magnetism" in 1879, he didn't *believe* a paragraph which started with the following statement:

"It must be carefully remembered, that the mechanical force which urges a conductor carrying a current across the lines of magnetic force, acts, not on the electric current, but on the conductor which carries it."[36]

Since electron beams are deflected by electric and magnetic fields, Maxwell's statement is evidently false. However, in 1879 electron beams had not been produced, so Maxwell's statement was in accordance with the experimental evidence of the time.

Hall expected that the current would be drawn to one side of the wire and that therefore the DC resistance would increase in the presence of the magnetic field. He found no evidence of this, even with his stated sensitivity being at the ppm level. He persevered with a thin foil of gold, looking for a potential difference perpendicular to the current flow and found one.

"In short, the phenomena observed were just such as you should expect to see if the electric current were pressed, but not moved, toward one side of the conductor."[37]

With a rectangular slab of conductor, it is found that when a magnetic field is applied perpendicular to the current, a potential difference is developed perpendicular to both the current and the magnetic field. This potential difference is found to be proportional to the current and to the magnetic flux density, but inversely proportional to the material thickness in the direction of the magnetic field.

The constant of proportionality is *defined* as the Hall coefficient.

$$\boxed{V_H = \frac{R_H \cdot I_X \cdot B_Z}{d}}$$

V_H is the Hall voltage, V, in the *Y* direction.

R_H is the Hall coefficient, m^3/C, also expressible as $V \cdot m/T \cdot A$.

I_X is the current, A, in the *X* direction.

B_Z is the magnetic flux density, T, in the *Z* direction.

d is the material depth (thickness), m, in the magnetic field (*X*) direction.

This relationship can also be expressed compactly as a vector product:

$$\boxed{\mathbf{E}_H = R_H \cdot \mathbf{J} \times \mathbf{B}}$$

$\mathbf{E}_H$ is the Hall electric field intensity vector, V/m.

$\mathbf{J}$ is the current density vector, A/m^2.

$\mathbf{B}$ is the magnetic flux density vector, T.

A charged particle moving through a magnetic field, whilst in the presence of an electric field, experiences a total force given by the vector equation $\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$, known as the Lorentz force. The cross product of the velocity and the magnetic flux density vectors gives a force perpendicular to both, causing the charge carriers to move over to one side of the conductor. The resulting increased charge density gives rise to an electric field, the Hall field. A dynamic equilibrium is therefore established with $\mathbf{E}_H = -\mathbf{v} \times \mathbf{B}$.

From this equality it is clear that a higher drift velocity gives a higher Hall field. The drift velocity will ideally be proportional to the current, but inversely proportional to the number of free charge carriers per unit volume. Thus highly conductive materials give low Hall coefficients.

The moving charge carriers can be of either polarity, and yet their transverse motion, in response to the applied magnetic field, is the same. For a given direction of current, both the velocity of the charge carriers and the relative direction of the force reverse when the polarity of the charge carriers

[36] J.C. Maxwell, 'Electromagnetic Force', in *A Treatise on Electricity and Magnetism*, 3rd edn (Clarendon Press, 1891; repr. Dover Publications, 1954), pp. Para 501, Vol 2.

[37] E.H. Hall, 'On a New Action of the Magnet on Electric Currents', in *American Journal of Mathematics*, II (1879), pp. 287-292.

is reversed. This reverses the polarity of the Hall field. Thus both the magnitude and the sign of the Hall coefficient give information about the nature of the conduction process in the material being tested.

Simplistically the Hall coefficient is $R_H = \frac{1}{n \cdot q}$, where *n* is the number density of the free charge carriers and *q* is their charge; *q* has both magnitude and sign. The estimate of *n* using the atomic density of a material multiplied by its valency is often wrong by orders of magnitude.

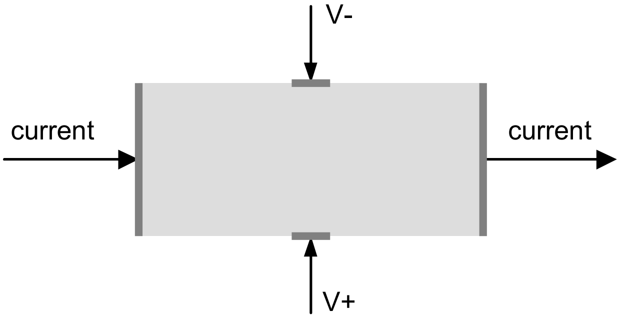
The Hall coefficients of common metals are of no great practical importance in themselves. However, the Hall effect was a breakthrough in the understanding of the internal workings of materials. It also brought to light a whole area of galvanomagnetic and thermomagnetic effects.[38]

Material	Hall Coefficient $\left(\frac{nV \cdot m}{T \cdot A}\right)$
Bismuth (Bi)	–600
Silver (Ag)	–0.08
Gold (Au)	–0.07
Copper (Cu)	–0.05
Aluminium (Al)	–0.04
Iron (Fe)	+0.8
Antimony (Sb)	+12
Silicon (Si)	+4000
Tellurium (Te)	+50,000

This table is presented to show the order of magnitude of the Hall effect. It has to be carefully understood that these values vary with the nature of the sample used, the temperature and the flux density applied. For example, the Hall coefficient of silver has been measured as –0.06 units at a film thickness of 0.12 μm, but double this at 0.05 μm.

The measured Hall coefficient is also strongly affected by direction in anisotropic materials. For example, Bismuth crystal cut in different planes gives more than 10:1 variation in the measured Hall coefficient.

Consider 1 A flowing in a copper track, 0.1 mm thick, through which a 1 T magnetic field is passing. The Hall voltage is only 0.5 μV; not very important.



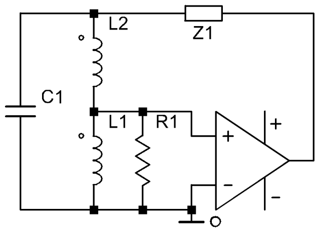
It is important to realise that the shape of the rectangular block affects the observed Hall effect. In this diagram, the magnetic field lines are perpendicular to the page. Notice that the current enters the specimen through electrodes which cover that whole face of the material. These electrodes therefore locally 'short out' the Hall voltage. The current could be supplied through smaller electrodes, but then there is the spreading of the current density to take into account. In both cases the situation is resolved by making the block longer in the direction of the current flow as shown in the diagram. If the block section is square then the voltage measured is around 30% lower than the simple equation predicts. A $length/width$ ratio of 3:1 is recommended for errors below 2%, with 4:1 being preferable. A correction factor could be used for short samples, but the uncertainty is necessarily greater due to the unknown uniformity of the end contacts.

HALF-WAVE LINE or PLATE: If a section of transmission line is not matched to the rest of the system, reflections will occur at both the input and output ends. However, if the line is half a wavelength long the reflections at each end of the line cancel making the line section 'transparent'. Consider a 50 Ω transmission system. Using a section of 100 Ω cable would ordinarily give reflections and despite this line section being lossless, the signal at the load is nevertheless reduced. At specific frequencies, $f = \frac{n}{2 \cdot T_P}$, the signal passes through without attenuation. *n* is an integer and T_P is the propagation delay down the cable.

[38] L.L. Campbell, *Galvanomagnetic and Thermomagnetic Effects: The Hall and Allied Phenomena.* (Longmans, Green and Co., 1923; repr., Johnson reprint Corporation, 1960).

Half-wave plates or films are also useful for either transparency or filtering in optical and mm-wave systems. In this case the film either reflects or transmits the incoming wave by 'selective interference'.

HARTLEY OSCILLATOR … is a sinusoidal LC oscillator with a tapped inductor chain which originated in 1915.[39] In this generic representation, the resonant circuit is formed by the series inductance of L1 & L2, in parallel with the capacitance of C1. R1 represents the finite input resistance of the amplifier and Z1 is a high impedance which includes the losses in C1 and L2.



If the amplifier has a current output then this circuit is a Thévenin equivalent, and the output of the amplifier (before Z1) will not be an accessible node. The output should not be taken directly from the resonant circuit because it will unduly reduce the Q. Hence it is usual to take the overall output from the bottom of the tapped inductor, which is at the *input* to the amplifier. It is usual for L1 to be much smaller than L2, minimising the loading by R1 on the resonant circuit, and maximising the Q.

HI-POT TESTING … is short for *High-Potential Testing*, also known as *Flash Testing*. The idea is to apply a high voltage to a component to check the integrity of the insulation. This sort of testing is routinely done on mains transformers, inlet filters and mains wiring. It is obviously important from a safety point of view. Whilst the hi-pot test can be done with DC, AC or transient voltages, a common theme is to ramp the voltage up over a period of tens of seconds; leave it at this full level for tens of seconds; then ramp it back down again at a similar rate to the initial increase. This slow ramp technique prevents extreme surge currents at switch-on and gives any sort of leakage path a chance to develop. Faulty components can tend to "flash over" so the hi-pot tester should have a built-in current limit to prevent explosive power levels.

Even components designed to run at only a few tens of volts may be flash tested at several hundred volts. The reason is to detect faults in the insulation. Such faults cause in-service failure, so flash testing can be done to detect flaws in the manufacturing process, thereby improving reliability. As a real life example, flash testing the coil to screen insulation on a 5 V reed relay using 100 V was found to be very successful at eliminating in-service failures (coil-to-screen short circuits).

IDC: Insulation **D**isplacement **C**onnector … a very inexpensive way of making a connection to a ribbon cable. The connector is pressed over the cable causing individual forked terminals to slice through the insulation and bite into the copper conductors. This is a very reliable and mature technology, but it goes horribly wrong if you:

☹ Use the wrong sized wire for the connector type you are using.
☹ Re-use the connector.
☹ Fail to use the correct tool to make the press-fit connection.
☹ Use the tool incorrectly, such as not pushing the wire in far enough.

IF: Intermediate **F**requency. In a *super-heterodyne* system, the incoming RF signal is *mixed* with a local oscillator (LO) having a similar frequency. The difference frequency is the Intermediate Frequency. When tuning a radio, the frequency of the local oscillator is changed and the IF is kept constant. The IF amplifier is tuned for optimum gain over a narrow range of frequencies. It is much easier to amplify signals at the IF rather than the incoming RF; gain is always more difficult to achieve at higher frequencies. Spectrum analysers use the same mixing and down-converting techniques. It has been normal on super-het systems to have several IF stages, with the operating frequency of each stage being 5× to 25× lower than the previous stage.

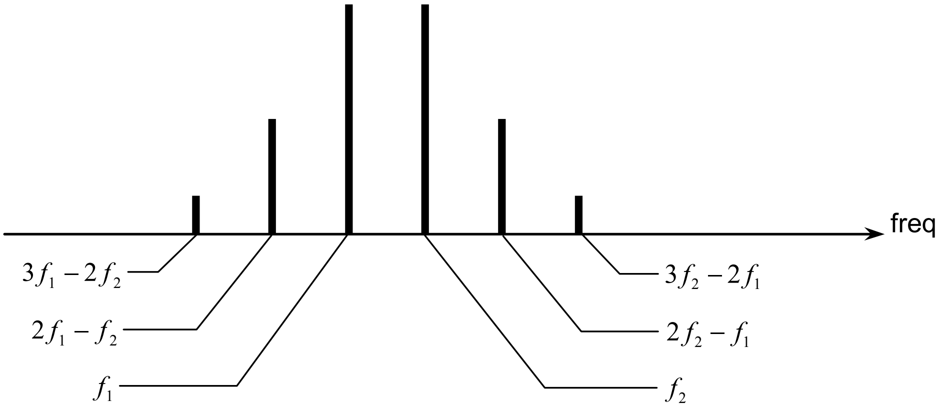
Modern designs use more stable oscillators and less IF stages. The direct conversion receiver, where an ADC samples the incoming RF signal, is becoming popular. The digital system is much more flexible than its analog counterpart, enabling channel switching to be done more rapidly.

[39] R.V.L. Hartley, 'Electric Oscillation Generators', *UK Patent Specification 141,046* (USPO, 1915: UKPO, 1921).

INTERMODULATION DISTORTION: When two different frequencies are simultaneously applied to a non-linear device, some degree of "mixing" takes place; sum and difference frequencies will be produced along with the original signals. A truly linear device will not produce this effect. Thus a quantitative measure of non-linearity of a device is the relative amplitude of the spurious frequencies generated when two equal amplitude sinusoids are applied to a device. This is known as a *two-tone intermodulation distortion test*.

In general the intermodulation products of f_1 and f_2 are $frequencies = \left| n \cdot f_1 \pm m \cdot f_2 \right|$, where *n* and *m* are both positive integers. The *order* of the intermodulation product is simply $n + m$.

If the input frequencies are fairly close together, the second order intermodulation products $(f_1 - f_2)$ and $(f_1 + f_2)$ will be a long way away from f_1. If the system is only operating over a narrow bandwidth, the second order intermodulation products will be out of band and can easily be filtered out. However, the odd-order difference-frequency intermodulation products may be in-band and may cause problems. (See www.logbook.freeserve.co.uk for software to calculate possible frequencies.)



Notice that all the odd-order difference-frequency terms are equally spaced from each other.

$$\Delta f = f_2 - f_1$$

$$2f_1 - f_2 = f_1 - (f_2 - f_1) = f_1 - \Delta f \qquad 2f_2 - f_1 = f_2 + (f_2 - f_1) = f_2 + \Delta f$$

$$3f_1 - 2f_2 = f_1 - 2(f_2 - f_1) = f_1 - 2 \cdot \Delta f \qquad 3f_2 - 2f_1 = f_2 + 2(f_2 - f_1) = f_2 + 2 \cdot \Delta f$$

Non-linearity in narrow-band RF systems is not characterised by harmonic distortion because the harmonics are outside the system passband. Non-linearity in narrow-band RF systems is therefore characterised by the odd-order difference-frequency intermodulation products. All the other intermodulation products will be out of band.

Because harmonically pure sources are expensive, it is sometimes preferable to infer harmonic distortion performance from two-tone intermodulation distortion testing. The two-tone intermodulation test does not require such harmonically pure sources. For example, rather than measure the second harmonic distortion directly, one can look at the $(f_1 + f_2)$ intermodulation product. This tone will be close in frequency to the second harmonic thereby minimising the flatness errors of the system.

The intermodulation product $(f_1 + f_2)$ will have the same amplitude as the second harmonic distortion that would be produced if the sources were harmonically pure, provided three key points are met:

1) intermodulation only occurs in the system under test and not in the generators themselves. This is ensured by the use of ***isolators*** and/or ***pads***.
2) Both generators are set to the same amplitude.
3) Only the second harmonic distortion due to the second order distortion term is considered.

The prediction of harmonic distortion in a real system does not take into account the higher order distortion terms. Therefore large input signals may generate more harmonic distortion that that calculated by this method.

ISOLATOR: A non-linear, non-reciprocal RF or microwave/mm-wave device, typically for use above 100 MHz, that has a low insertion loss in the forward direction, ≈1 dB, but has a return loss greater than ≈16 dB. An isolator is often used at poorly matched inputs or outputs to get a good Z_0-match without sacrificing signal strength.

Junction isolators are narrow band, giving <1% fractional bandwidth, whereas Faraday rotation devices can give full band operation in waveguide. More than an octave of bandwidth is unusual. Isolators based on ferrite and permanent magnets are very sensitive to external magnetic fields.

JITTER … refers to a random or systematic change in the delay of a signal through some device or system. In logic systems, the effective propagation delay will not be constant. The jitter is due to noise on the switching thresholds and is therefore made worse by slow input edge-speeds and slow internal edge-speeds. If this jitter were large enough it could be viewed on a scope by triggering on the input signal and viewing the output signal, possibly using a *delayed trigger*. The signal will always jitter; being able to measure this jitter depends only on the resolution and jitter of the measuring system.

Phase noise on the master clock of a logic system gives jitter on the logic edges. The phase noise is a *frequency domain* measure and the jitter is a *time domain* measure.

Inexpensive (<$0.50) logic gates do not have specified jitter. As far as the logic gate spec is concerned, the output can occur at any time between the max and min propagation delay limits. You should expect jitter to be an increasing function of propagation delay, and both internal and external edge-speeds. A gate with a long propagation delay will almost certainly jitter more than one with a short propagation delay.

Direct measurements of jitter in the pico-second region are exceptionally difficult. Often measurements are made indirectly and the result inferred. One method uses the reduction in signal-to-noise ratio produced on an ADC clock when sampling a pure sinusoidal signal. Jitters on simple logic gates have been measured on this basis:[†] 74LS00= 5 ps; 74HCT00= 2 ps; 74ACT00= 1 ps.

All indirect measurement techniques suffer from the same problem: the calculation of the jitter requires a conversion from the measurement to the answer. This conversion is often based on an inadequate model of the nature of the jitter. It has to be understood that jitter can have both random and deterministic elements. The random elements are often assumed to be Gaussian but may also have 1/f characteristics. The deterministic elements may be periodic, data dependant, power supply related and so forth. Calculations based on Gaussian jitter can therefore give wildly inaccurate results when the jitter is actually dominantly deterministic.

In any real system using single-ended logic (as opposed to differential logic such as ECL, PECL and LVDS) the ground and power rails transmit part of the signal. Thus the power/ground noise will increase the jitter, depending on how well or badly the circuit is layed out, and how far apart the gates are.

Manufacturers of crystal oscillator modules often quote RMS jitter, particular on high-end differential PECL output devices, and figures between 1 ps and 10 ps are not unusual. Low-jitter applications should always use a differential clock whenever possible.

See also ***aperture jitter***.

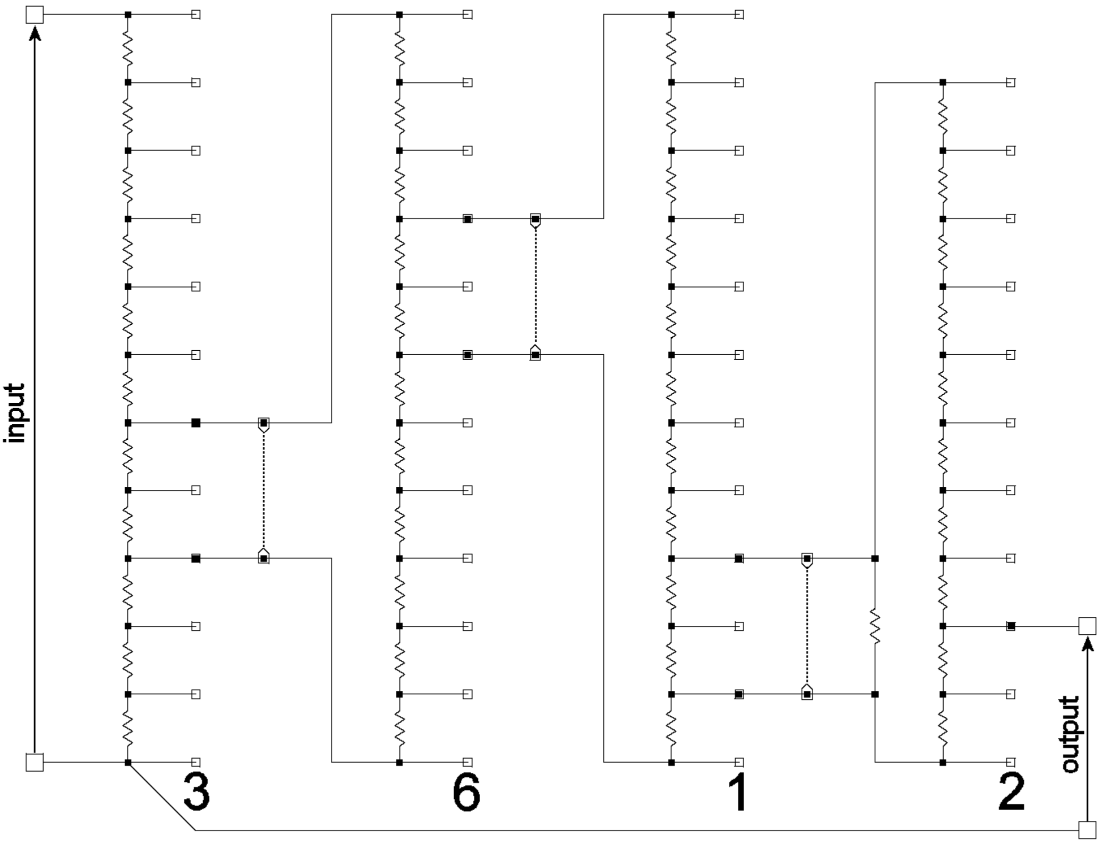
KELVIN … is the name of the British physicist Sir William Thomson FRS (1824-1907). He became Lord Kelvin in 1892. It is also the name of the absolute temperature scale named in his honour.

KELVIN CLIPS … a set of test leads with two pairs of force-sense connections. It is possible to buy crocodile clips with the two halves of the clip electrically isolated from each other. One is used as a ‘force’ [eg I+] and the other a ‘sense’ [eg Hi], enabling a true 4-wire connection to be made to a

[†] B. Brannon, *Aperture Uncertainty and ADC performance*; Analog Devices application note AN-501 (rev 0.)

component. This scheme eliminates the problem of the contact resistance of the connections, but does not eliminate all measurement uncertainties. The *current spreading* from the 'force' contacts will mean that the resistance reading will necessarily be variable according to the exact position of the clips. This uncertainty is reduced by clipping onto a relatively large contact area, which then thins down to the relatively small part of the resistive element.

KELVIN-VARLEY DIVIDER … is an ingenious cascadable passive resistive divider used to make ultra-precision potentiometric voltage comparisons.[40] By using equal valued resistors, which can be calibrated against each other, division ratios with a linearity of around ±0.2 ppm can be made.[41] This used to be the primary standard for DC linearity calibrations. It has now been replaced by superconducting Josephson junction arrays for the very highest precisions available.



This 4 decade Kelvin-Varley divider is set to a ratio of 0.3612 of the input. The divider is switched with 2-pole 10-way switches, the dotted bars indicating how the 2 poles are mechanically ganged together. Notice that the attenuator decades have 11 resistors each and that 2 resistors in the chain are shunted by the next stage. Successive stages have resistor values a factor of 5 lower than the previous stage. The pair of resistors at the output of each stage is therefore loaded by an equal resistance, making the combination equal to a single resistor in that chain. Notice that the last stage needs one less resistor in the chain and only a single-pole switch.

The resistors need to have low self-heating errors, since the output pair in each stage run at half the current, and therefore one quarter the power, compared to the other resistors in that chain.

On a 7-decade divider, the resistors in the final stage would be $5^6 = 15625\times$ smaller than the resistors in the input stage, making the resistors and the switching less accurate. This problem can

[40] C.F. Varley, 'On a New Method of Testing Electric Resistance', in *Report of the British Association for the Advancement of Science*, [Notices and Abstracts Section] (1866), pp. 14-15.

[41] A.F. Dunn, 'Calibration of a Kelvin-Varley Voltage Divider', in *IEEE Transactions on Instrumentation and Measurement*, IM-13 (1964), pp. 129-134.

be overcome by using additional precision shunt resistors across the inputs to the lower decade stages. One such resistor has been shown in the figure at the input to the last stage.

LAPLACE's EQUATION: The French mathematician Pierre-Simon, marquis de Laplace, discovered this equation around 1785. For a three-dimensional field using rectangular coordinates:

$$\boxed{\frac{\partial^2 V}{\partial x^2}+\frac{\partial^2 V}{\partial y^2}+\frac{\partial^2 V}{\partial z^2}=0}$$

This second order partial differential equation of potential is very important for solving electric, magnetic, thermal and other field problems. This complicated looking equation does not represent any difficult concepts, however.

Consider water flowing in a shallow, smooth flowing river. If you were to sketch a plan view of the river and you were to draw the outline of any closed shape on the plan, you would readily agree that the amount of water flowing into this closed shape was equal to the amount of water flowing out of the closed shape. If not, there would be an accumulation or depletion of water in this region. Water is neither being created nor destroyed, so the amount going in is equal to the amount coming out. Obviously evaporation, and drainage through the water bed, are being conveniently neglected.

In more general terms, think about the *flux* of a quantity, flux meaning flow. This might be a heat flux (measured in watts), a current flux (measured in amps), a magnetic flux (measured in Webers), or the flux of an electric field (measured in Coulombs). Provided these quantities are not being created within the region, Laplace's equation applies.

In situations where a flux is being created within the region, *Poisson's equation* is used. In a resistive film with significant power dissipation, the heat flux out of a region is equal to the heat flux into the region *plus* the heat flux generated within that region. These are not complex concepts when applied to physically understandable fields. They just become difficult to visualise when applied to more 'abstract' quantities such as electric flux. It is therefore useful to think in terms of one of the analogous situations such as water flow or heat flow.

Both magnetic flux and electric flux are quite *real* in the sense that you can measure them. However, one would not really consider the electric flux or magnetic flux to actually be moving. Take the case of a permanent magnet producing a static magnetic field. If one considers the magnetic flux to be moving then this seems like a perpetual motion without the influx of energy, a concept abhorred by scientists and engineers alike.

In two-dimensional cylindrical coordinates, Laplace's equation becomes:

$$\boxed{\frac{\partial^2 V}{\partial r^2}+\frac{1}{r}\frac{\partial V}{\partial r}+\frac{1}{r^2}\frac{\partial^2 V}{\partial \phi^2}=0}$$

Notice that Poisson's equation is not used with magnetic fields because magnetic monopoles are not a naturally occurring phenomenon, have not yet been created artificially, and may not even exist! For regions containing unbalanced charges, however, it is essential to use Poisson's equation when solving the field distribution.

LANGE COUPLER … is a 4-port narrow-band microwave device based on inter-digitated quarter-wave microstrip lines; alternate microstrip lines being linked together using bond-wires.[42] One application is to split a single-ended signal into a pair of 90° separated (quadrature) signals to feed into an amplifier. Another coupler would then be used at the output to recombine the signals. Since this coupler is based on $\lambda/4$ microstrip lines, it becomes unwieldy below 300 MHz.

LEAST SQUARES REGRESSION: The *least squares fit* to a set of data points is a useful method of computing a trend-line without any unintentional bias from the experimenter. The least squares line gives the minimum RMS value for the difference between the trend-line and the data; in this sense it gives the least 'noise'. Other ways of fitting a curve to data points include: a) Mini-max fit, where the maximum deviation from the trend-line is minimised; b) Mean fit, where the mean deviation from the trend-line is zero. This means that the data points are equally balanced either side of the trend-line.

[42] J. Lange, 'Interdigitated Stripline Quadrature Hybrid', in *IEEE Transactions on Microwave Theory and Techniques*, MTT-17, no. 12 (Dec 1969), pp. 1150-1151.

Least squares polynomials can be calculated directly from experimental data, without needing an iterative process. Unfortunately the formulae are not generally written out conveniently for use.

For a best fit straight line through the data points, the method is known as *linear regression*. The best fit (regression) line is $y = m \cdot x + c$, the basic equation of a straight line. The required values are the slope, *m*, and the offset, *c*.

The experimental data will be in the form of *N* data pairs of *x* and *y*, from which the slope and offset are calculated.

$$\boxed{m = \frac{\left(\sum_{r=0}^{N-1} x_r y_r\right) - \frac{1}{N}\left(\sum_{r=0}^{N-1} x_r\right)\left(\sum_{r=0}^{N-1} y_r\right)}{\left(\sum_{r=0}^{N-1} x_r^2\right) - \frac{1}{N}\left(\sum_{r=0}^{N-1} x_r\right)^2}}$$

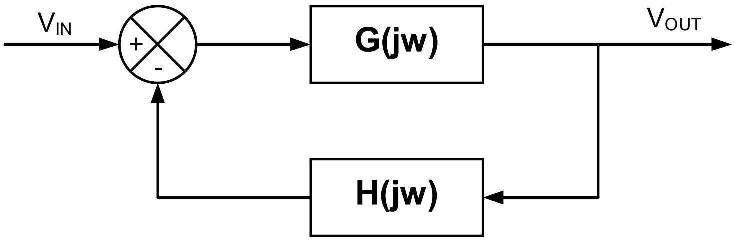
$$\boxed{c = \frac{1}{N}\left(\sum_{r=0}^{N-1} y_r\right) - \frac{m}{N}\left(\sum_{r=0}^{N-1} x_r\right)}$$

Often the *x* values are acquired at a fixed interval, from the timebase of a sampling system for example. In this case $x_r = r$, giving the simplified forms:

$$\boxed{m = \frac{12}{N\left(N^2 - 1\right)} \cdot \left[\left(\sum_{r=0}^{N-1} x_r y_r\right) - \frac{N-1}{2} \cdot \sum_{r=0}^{N-1} y_r\right]}$$

$$\boxed{c = \frac{1}{N}\left(\sum_{r=0}^{N-1} y_r\right) - \frac{m}{2}\left(N - 1\right)}$$

LOOP-GAIN:



In control system theory, a control system is drawn as a forward path G(jω), a feedback path H(jω) and a perfect subtractor block.

This has the transfer function

$$\frac{V_{OUT}(j\omega)}{V_{IN}(j\omega)} = \frac{G(j\omega)}{1 + G(j\omega) \cdot H(j\omega)}$$

This equation ‘blows up’ when the term $G(j\omega) \cdot H(j\omega)$ in the denominator becomes –1. In fact the nature of this product around the –1 point is so critical that the equation $1 + G(j\omega) \cdot H(j\omega) = 0$ is known as the *characteristic equation* of the system. The product $G(j\omega) \cdot H(j\omega)$ is called the *loop-gain*.

Bode plots and Nyquist plots are both graphs of the loop-gain against frequency. Note that the term ‘loop-gain’ represents the complex gain; it has both magnitude and phase. The *open-loop gain* is $G(j\omega)$ and tells you nothing about the stability of the system.

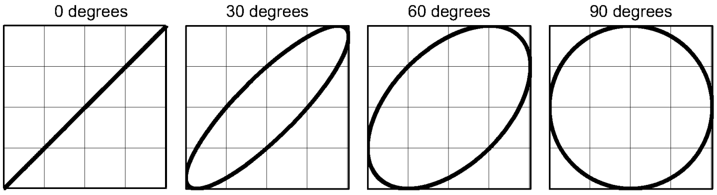
In a stable system, as frequency increases, the loop-gain magnitude falls below unity before the loop-phase shift reaches 180° (π radians). The amount by which the gain is lower than 1 is the ***gain margin***; usually expressed in dB.

$$\boxed{\begin{array}{l} \text{Gain Margin} = -20 \cdot \log_{10}\left|G(j\omega) \cdot H(j\omega)\right| \text{dB} \\ \text{when } \arg\left[G(j\omega) \cdot H(j\omega)\right] = \pi \text{ radians} \quad \text{for the first time, as } \omega \text{ increases from zero.} \end{array}}$$

In a stable system the loop phase shift must be well away from 180° when the loop-gain magnitude becomes 1; this difference is the ***phase margin***.

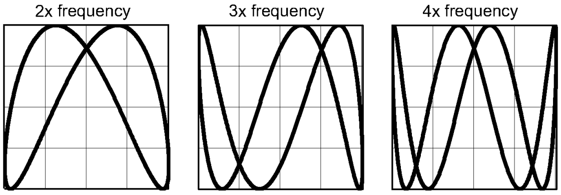
$$\boxed{\begin{array}{l}\text{Phase Margin} = \pi - \arg\left[G(j\omega)\cdot H(j\omega)\right]\text{radians}\\ \text{when}\left|G(j\omega)\cdot H(j\omega)\right| = 1 \quad \text{for the first time as } \omega \text{ increases from zero.}\end{array}}$$

LISSAJOUS FIGURES … are named after the French mathematician J.A.Lissajous, who investigated the curves in detail around 1857. He used a narrow stream of sand flowing from a container on the base of a compound pendulum to produce the patterns. It is inconvenient to try to pronounce this as Lissajous' figures, so the possessive apostrophe is best left off.



For sinusoidal signals of similar frequency, compared using the X-Y display of a scope, the Lissajous figure appears to be a circle rotating at the frequency difference between the two signals. For exactly equal frequencies, the relative phase difference is the arcsin of the ratio of the vertical 'open' height to the overall height of the figure. You can see this on the 30° figure, where the open width at the centre is 2 divisions and the overall height is 4 divisions: $\arcsin\left(\frac{2}{4}\right) = 30°$ phase difference.

$$\boxed{\text{phase difference} = \arcsin\left[\frac{\text{open height}}{\text{maximum height}}\right]}$$



More complicated Lissajous figures are seen when the sinusoidal signals are harmonically related to each other.

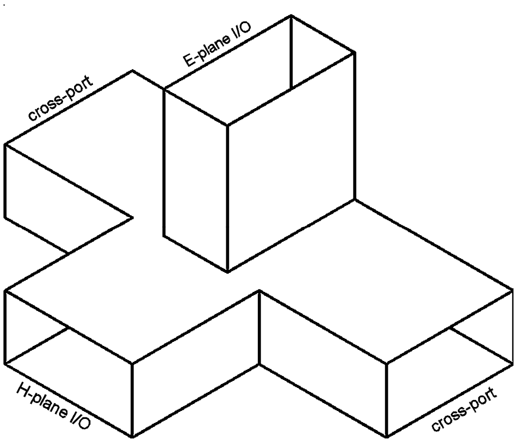
If the sinusoids have a suitable phase shift, you can count the peaks of the figure to determine the frequency ratio. Again, if the frequencies are not exactly locked, the patterns will seem to rotate.

LITZ WIRE: Litz is short for the German word *Litzendraht* meaning *braided wire*. The ***skin effect*** means that alternating current always flows on the outermost surface of a conductor. When the skin depth is smaller than the conductor diameter, the resistive loss increases proportionately to the square root of frequency. To reduce this loss, the conductor can be made of insulated strands, woven together; the finished cable being known as *Litz wire*. Each strand is shifted from the inner part of the bundle to the outer part of the bundle as one proceeds down the length of the cable. This weaving allows the current to be shared equally amongst the conductors and minimises the AC resistance of the composite wire. This constructional technique is only workable up to around 1 MHz. Above a few megahertz, Litz wire is more lossy than solid wire. Because of the lower resistance, inductors for radio receivers in the frequency bands below a few hundred kilohertz have a higher Q when made of Litz wire.[43] Litz wire was commercially manufactured at least as early as 1898.

MAINS: The AC power from a local supplier, or a National grid, that you would plug electrical appliances into at work or in the home. This term is widely used in the UK, Hong Kong, Malaysia, Singapore and a few Middle Eastern countries. It is also used in international standards (eg EN60950-1:2006 clause 1.2.8).

MAGIC-T … is a passive four-port waveguide junction for microwaves, also known as a hybrid-junction or a hybrid-T. All four ports are connected to rectangular waveguides. Applications include impedance bridges, balanced mixers and balanced *duplexers*.

[43] B. Bowers, 'Low Frequency Coil 'Q', in *The LF Experimenter's Source Book*, ed. by Dodd, P., 2nd edn (Radio Society of Great Britain, 1998), pp. 1.21-1.23.



There is ideally minimal straight-through coupling between the input & output ports. In the drawing the connecting flanges on the I/O ports have been omitted for simplicity.

The top I/O port is known as the E-plane port, the cross-port axis being parallel to the **E**-field in that port. A signal fed into the E-plane port results in cross-port outputs which are in anti-phase with each other. The lower I/O port is known as the **H**-plane port. A signal fed in this port also splits equally, but both signals are in-phase with each other.

The magic-T requires internal metallic structure, such as a tuning rod, in order to improve the matching at the I/O ports.

In one application, a signal from a microwave oscillator is fed into the magic-T input. The signal splits equally at the junction and heads from the cross-ports towards the loads. If these loads are matched to the waveguide impedance then there is no reflection and no signal arrives back at the detector. The attenuation directly from input to output of the magic-T might be 40 dB. In general, the signal at the output is ideally the difference between the signals entering at the cross-ports. This system is therefore a microwave VSWR bridge.

In another application a microwave transmitter and receiver are connected to the I/O ports, with an antenna on one of the cross-ports, the other cross-port being Z_0 terminated. This could form the basis of a radar system.

MARK/SPACE RATIO: The mark is the “on time” and the space is the “off time”. It is more common in modern usage to find ***duty cycle*** given instead of mark/space ratio.

$$\boxed{Duty\ Cycle = \frac{on\ \ time}{cycle\ \ time} = \frac{Mark}{Mark + Space}}$$

MATCHED SYSTEM: At DC (and LF), maximum power transfer occurs when the load resistance equals the source resistance. Note that this scheme does not give the minimum voltage attenuation however. The minimum voltage attenuation is achieved for an infinite input impedance.

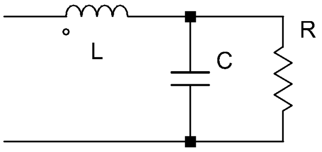
For RF systems (at any frequency), matching for maximum power transfer is called *conjugate-matching*. If the source impedance is R + jX, the load impedance is made the *complex conjugate* of the source impedance, namely R – jX. Care must be taken when using or reading the unqualified term “matched”, since it is also used for the minimum reflection condition, a *Z_0-match*. It is best to be explicit and state which match is meant. When reading the literature, the term ‘matched’ often means a Z_0-match, maximum power transfer matching being explicitly referred to as a conjugate match.

Another definition of matching is when two or more components have properties that are selected or adjusted to be similar. Thus a pair of resistors can be called a ‘matched pair’ if they are chosen for similar TCs or similar nominal values. Matching achieves a better system response without requiring increased absolute accuracy from the individual components.

MATCHING NETWORKS: Optimum power transfer requires load and source resistances to be equal. When the load and source resistances are mismatched, a network can be used to make the load resistance equal to the source resistance. In elementary courses this idea is demonstrated using matching transformers. If there is no need for isolation, simpler networks are possible. It is also difficult to produce conventional transformers above 100 MHz. One possibility for operation above 100 MHz is the use of a ***quarter-wave transformer***. This solution is not feasible at lower frequencies because the transmission line would be inconveniently long. *Lumped component* solutions are possible at any frequency.

If the load resistance is too high, a capacitor placed across it will lower the real part of the resulting impedance. This only becomes clear when you follow the equations:

$$Z=\frac{1}{\frac{1}{R}+j\omega C}=\frac{R}{1+j\omega CR}=\frac{R}{1+j\omega CR}\cdot\frac{1-j\omega CR}{1-j\omega CR}=\frac{R}{1+(\omega CR)^2}-j\frac{\omega CR^2}{1+(\omega CR)^2}$$



The resulting impedance is seen to be equivalent to a reduced resistive component in series with an unwanted capacitive reactance. The capacitive reactance can be cancelled by the addition of an equal series inductive reactance. The resulting network is called *an L-section*, not because of the inductor, but because of the shape of the network.

The resistance has been reduced by a factor $k=1+(\omega CR)^2$.

The required capacitor value is therefore …

$$\boxed{C=\frac{\sqrt{k-1}}{2\pi f\,R_{LOAD}}}$$

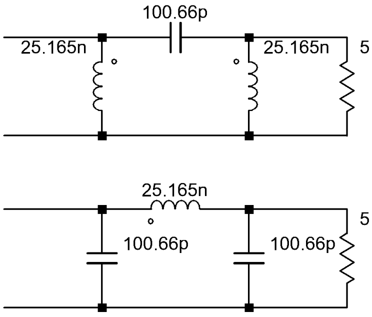
The inductive reactance must equal the capacitive reactance in *Z*.

$$\omega L=\frac{\omega CR^2}{1+(\omega CR)^2}=\frac{\omega CR\cdot R}{k}=\frac{\sqrt{k-1}}{k}\cdot R$$

$$\boxed{L=\frac{R_{LOAD}}{2\pi f}\cdot\frac{\sqrt{k-1}}{k}}$$

Remember that *k* is the desired transformation ratio, $\boxed{k=\frac{R_{LOAD}}{R_{SOURCE}}}$

If the load already contains some shunt capacitance, the value of *C* can be reduced to allow for it. Likewise, if the source already contains some series inductive reactance, the value of *L* can be reduced accordingly. This technique is known as *absorption* of the ***parasitics*** and strays. The inductor and capacitor can be interchanged to suit the bias conditions, the values remaining unchanged.



If the load resistance is lower than the source resistance, the *L*-section network will not work. One possibility is to use a symmetrical PI network designed to be the lumped equivalent of a quarter-wave line.[44]

The two possible types of symmetrical PI networks allow a choice based on bias conditions. These example networks match a 5 Ω load to a 50 Ω source at 100 MHz.

The reactance of each element in the PI network is the geometric mean of the source impedance and the load impedance.

$$\boxed{L=\frac{1}{2\pi f}\sqrt{R_{LOAD}\cdot R_{SOURCE}}=\frac{R_{LOAD}}{2\pi f\sqrt{k}}}\qquad\boxed{C=\frac{1}{2\pi f\sqrt{R_{LOAD}\cdot R_{SOURCE}}}=\frac{\sqrt{k}}{2\pi f\,R_{LOAD}}}$$

A quarter-wave transformer always gives a larger bandwidth than the lumped component solution. For a 2:1 transformation at a VSWR limit of 1.2, the λ/4 transformer gives 2.2× the bandwidth. For a 10:1 transformation and a VSWR limit of 1.5, the bandwidth improvement is only 35%.

MEISSNER EFFECT: When a material is cooled sufficiently for it to become superconducting, its resistance *abruptly* changes to zero. It also becomes *perfectly diamagnetic*, meaning that it excludes any magnetic field from passing through it. A perfectly conducting material would be predicted to prevent any *change* in the magnetic field passing through itself. However, a material through which a magnetic field is passing will expel this field as the material enters the superconducting state; this is the Meissner Effect, discovered in 1933.

[44] R.P. Glover, 'R-F Impedance-Matching Networks', in *Electronics*, 9 (Jan 1936), pp. 29-30.

The magnetic field decays exponentially with distance into the superconductor, the decay being given by $H(x) = H(0) \cdot \exp\left(-\frac{x}{\delta}\right)$

where δ is the *penetration depth*. The penetration depth changes rapidly from an effectively infinite value to perhaps a few tens on nanometres at the critical temperature of the superconductor. Further reduction of the temperature decreases the penetration depth. At 0°K the penetration depth in Niobium is 47 nm, whilst that in Lead is 39 nm. Penetration depths of 200 nm at 0°K have been observed in high temperature superconductors.

Notice that whilst any good conductor will tend to exclude an alternating magnetic field, a superconductor excludes a steady ("DC") magnetic field as well.

METASTABLE STATE: Strictly speaking this is a digital problem, but when edge speeds are down to the sub-nanosecond region, an analog approach is necessary. Consider a D-type flip-flop. If the signal at the D-input is changing around the time that the active clock edge is changing, the ***set-up or hold times*** of the gate may be violated. In this case the output of the flip-flop is not defined.

You might suppose "not defined" means the output will either be a 0 or a 1, and that it is not clear which one it will be. Unfortunately this is not always the case. The gate can get into a *metastable* state, where its output is neither a 0 nor a 1. It can hover between the two valid logic levels, as if uncertain which way to go. In fact it can stay in this state for a relatively long time. For fast ECL devices this might be of the order of a nanosecond. The gate can also go high then fall back to a low as its metastable response. Asynchronous signals are therefore often clocked through two latches to ensure that the result is not metastable.

METROLOGY: The science of measurements and standards, derived from the Greek *métron* meaning measure, and the ending -*logy* meaning study or science of.

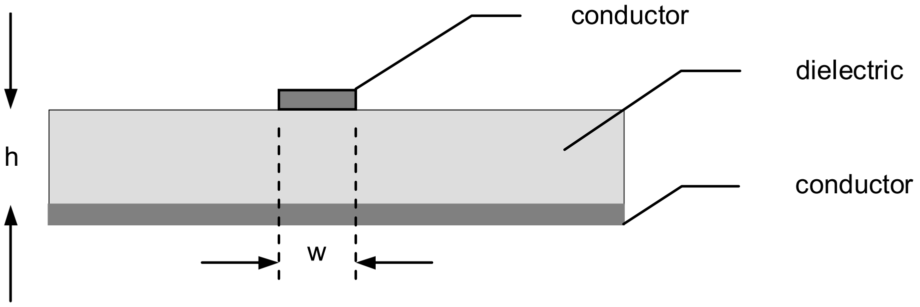
MICROSTRIP: If a PCB has a ground plane on one side and a track of defined width on the other side, this is described as *microstrip*. Such a track has a well defined characteristic impedance. When people talk about 'controlled track impedances' on PCBs, they are usually considering the tracks as microstrip. Microstrip originated as a low-loss microwave interconnection scheme.[45]

On a multi-layer board having at least one ground or power plane, any track can be considered as a microstrip line, provided that it is not sandwiched between ground or power planes. A track with ground/power planes on both sides of it is modelled as a *stripline*, even if the ground/power planes are not on the immediately adjacent layers.

For microstrip, the relevant factors are the spacing of the track from the ground plane, the dielectric constant of the *substrate* {PCB material}, the width of the track and how close other tracks are to it. If it is desirable to have a track with a controlled impedance, then it is unwise to put right angled bends in it. The worst bend should be 45°, a smooth curve being even better. This requirement is due to the fact that the bend will give a discontinuity in the characteristic impedance, and therefore a reflection.

[45] D.D. Grieg, and H.F. Engelmann, 'Microstrip - A New Transmission Technique for the Kilomegacycle Range', in *Proceedings of the Institute of Radio Engineers*, 40 (Dec 1952), pp. 1644-1650.

The minimum sideways spacing to other tracks should be greater than the largest of the track width, *w*, and the spacing, *h*, to the ground plane. Double this spacing is preferable, but much more than this would not ordinarily be necessary.



The ground plane should be larger than either 3×*w*, or 3×*h*, whichever is the greatest. In many cases, it is convenient to use the equations for microstrip and ***coplanar waveguide*** in order to estimate the performance of a printed circuit track layout.

Making the track wider increases the capacitance and therefore makes the characteristic impedance lower. The impedance equations are complicated, but well documented.[46]

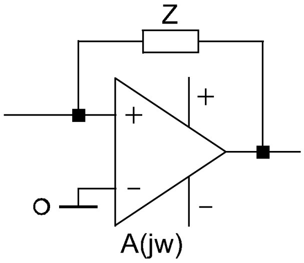
In order to avoid dispersion, caused by higher order modes,[47] it is important to limit the maximum operating frequency below $f_c = \frac{c_0}{(2w + 0.8h)\cdot\sqrt{\varepsilon_r}}$, where $c_0 = 3\times10^8$. Dispersion is therefore not normally a problem except on mm-wave circuits; $\lambda_c = (2w + 0.8h)\cdot\sqrt{\varepsilon_r}$

To inspect the impedance of a track, it is reasonable to apply a fast step to the track. If the tracks on either side now have the same step signal applied when the measurement is undertaken, the result will be very different to the cases where the adjacent tracks are grounded or left open-circuit. If the adjacent signals are in anti-phase the measurement will be significantly different again. The idea of a fixed characteristic impedance of the track is therefore not valid.

A typical example would be the case of a digital data bus. If the adjacent bits happen to be changing in the same direction as the bit-line in question, the measured impedance will be considerably different to that seen when the adjacent bits happen to be changing in the opposite direction. This cross-talk situation is considerably improved by the addition of ground tracks in amongst the data lines. Obviously having a ground, power, or relatively static control line in between each data bit is the ideal situation. In practice, it may be necessary to use one of these ground-like tracks only for every other bit, in order to reduce the overall width of the data bus.

MIL: US term, 1 mil ≡ 0.001 inch. In the UK, 1 thou ≡ 0.001 inch.

MILLER CAPACITANCE: Miller's theorem allows an impedance connected across an amplifier to be replaced by an impedance to ground.



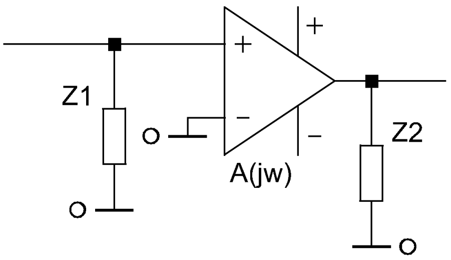
The amplifier has a frequency dependent voltage gain A(jw). If the voltage input is V_{IN}, the output will be $V_{IN} \times A(j\omega)$. The current in the impedance is then: $V_{IN} \times \frac{[1 - A(j\omega)]}{Z}$.

The impedance *Z* can therefore be replaced by equivalent impedances Z_1 and Z_2.

[46] M.A.R Gunston, Microwave Transmission Line Impedance Data (Van Nostrand Reinhold, 1972; repr., Noble Publishing Corp., 1996).

[47] N. Kinayman, and M.I. Aksun, '2.2.5 Higher-Order Modes and Dispersion' in Modern Microwave Circuits (USA: Artech House Inc., 2005), pp. 161-163.

$$Z_1 = \frac{Z}{1 - A(j\omega)} \; ; \; Z_2 = \frac{Z}{1 - \frac{1}{A(j\omega)}}$$



If the impedance is a capacitor and the amplifier is inverting, the effective capacitive loading is increased by the factor $1 + A(j\omega)$. This is *Miller capacitance*.

If the amplifier is non-inverting and $A(j\omega) \approx 1$, then the input impedance is increased; the impedance has been ***bootstrapped***.

MIXER: For audio systems, a mixer *linearly adds* signals together. For RF systems a mixer is designed to produce the maximum non-linearity, deliberately producing *harmonics* and *intermodulation products*. An RF mixer is more like a multiplier than an adder. For mm-wave operation it is exceptionally difficult and expensive to generate the local oscillator power. In this case it is helpful to use an LO at half the desired LO mixing frequency. Such a mixer is described as being *sub-harmonically pumped* or simply a sub-harmonic mixer for short. Typically the active part of such a mixer is simply a pair of gallium arsenide or indium phosphide schottky diodes connected in inverse-parallel. Filters are required to separate the LO and RF paths for optimum conversion loss.

MONOTONIC … means that the signal is always changing in the same direction. A frequency response curve that is always decreasing with increasing frequency is termed monotonic. Increasing quantities continue to increase; decreasing quantities continue to decrease. In calculus terms, the second derivative of the signal does not change sign.

See graph under ***non-monotonic***.

MTBF: **M**ean **T**ime **B**etween **F**ailures. This is a statistical measure of the reliability of a repairable component or a system.

$$\boxed{MTBF = \frac{\text{Total operating Time of all units in the field}}{\text{Number of Failures in that time}}}$$

It is common practice to estimate the MTBF by using individual failure rate data for the components. This is mostly restricted to military and aerospace activities.

From a quality control viewpoint, MTBF can be measured by looking at the number of units repaired with time against the number of units in service. This is a genuine statistic which can be used to assess the quality of a production process. It is not much use for quality control, however, because if there was a faulty batch three months ago then all those units may start failing now. There is nothing that can now be done to stop those units from failing. Thus the control aspect has to be done earlier in the failure cycle. The MTBF figure assures *future* customers that the previous quality levels have been high. Alternatively, for safety related equipment, it could warn the user of a need to take a piece of equipment out of service.

MU-METAL: A ferromagnetic material with an LF μ_r {relative permeability} which is remarkably high (>10,000). It is used to shield against low frequency (<3 kHz) magnetic fields. Such fields might come from AC power transformers, magnetic deflection coils for CRTs, high-current wiring, or even the Earth's magnetic field. Once a shield has been shaped, it has to be annealed {heat treated} to give it optimal magnetic performance. If it is dropped, hit or bent, its relative permeability can be dramatically worsened (up to say threefold reduction).

NEPER: The Scottish mathematician J.Napier (also spelt Neper) published the original work on *natural* logarithms in 1614. For this reason natural logarithms, that is logs to the base *e*, are also known as Naperian logarithms. The Neper is defined as a ratio of voltages or current that is equal to *e*. [e=2.718281]

$$\boxed{\text{voltage ratio in Nepers} \equiv \log_e\left(\frac{V_1}{V_2}\right) \equiv \ln\left(\frac{V_1}{V_2}\right)}$$

1 Neper = 8.6859 dB

The conversion to decibels is seen below:

$$1\,\text{Neper} = \ln\left(\frac{V_1}{V_2}\right) = \ln(e) \Leftrightarrow \quad 20\cdot\log_{10}\left(\frac{V_1}{V_2}\right) = 20\cdot\log_{10}(e) = 8.6859\,\text{dB}$$

Henry Briggs felt that logarithms to the base of 10 would be more useful and published his work in 1617. Hence common logarithms, that is logs to the base 10, have historically been referred to as Briggsian logarithms. Briggs published tables of common logs to 14 places in 1624.

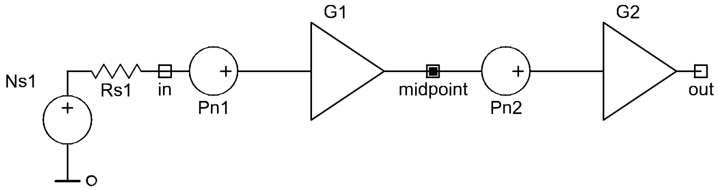
NOISE FACTOR: Modern usage make a distinction between *noise factor*, *F*, and *noise figure*, *NF*. The relationship is that noise figure is noise factor expressed in decibels. These two terms cannot be confused if you keep the units in mind, but can easily be confused when carelessly applying formulae from text books. Noise factor is a power ratio or an RMS voltage ratio squared.

$$F = \frac{\text{ideal input signal to noise (power) ratio}}{\text{available output signal to noise (power) ratio}} = \frac{P_{SIG}/P_{NOISE,IN}}{G\times P_{SIG}/P_{NOISE,OUT}} = \frac{P_{NOISE,OUT}}{G\times P_{NOISE,IN}}$$

G is the ***available power gain***. The output noise is the input noise multiplied by G, plus a contribution due to the device. Dividing the device output noise by the *power* gain, refers the device noise to the input.

$$\boxed{F = \frac{G\times P_{NOISE,\,IN} + P_{DEVICE,\,OUT}}{G\times P_{NOISE,\,IN}} = 1 + \frac{P_{DEVICE,\,IN}}{P_{NOISE,\,IN}}}$$

It is evident that F > 1. The input noise power is *defined* as the ***Johnson noise*** in the source resistance, held at 290°K. When stages are cascaded, it is convenient to think of the device noise as a noise (power) generator in series with the input. Noise power adds, the sources being considered as uncorrelated.



The output of the two cascaded stages shown, assuming infinite input impedances, is therefore:

$$P_{OUT} = \left[(N_{s1} + P_{n1})\cdot G_1 + P_{n2}\right]\cdot G_2 = \left(N_{s1} + P_{n1} + \frac{P_{n2}}{G_1}\right)\cdot G_1 G_2$$

This formula clearly shows that the noise of the second stage is attenuated by the gain of the first stage and is often therefore relatively unimportant.

The input referred noise power sources can easily be related to the noise factors of the devices. Since N_{s1} is the noise power in the source resistance R_{s1}, $F_1 = 1 + \frac{P_{n1}}{N_{s1}}$

In a 50 Ω system, all the inputs and outputs would be nominally 50 Ω, but in general there is no requirement for this. The noise factor of the second stage has to be measured and specified relative to the output impedance of the first stage if the two noise factors are to be combined simply and correctly. Assuming that the impedances are all the same:

$F_2 = 1 + \frac{P_{n2}}{N_{s1}}$ and the power output equation can be re-expressed in terms of noise factors.

$$P_{OUT} = \left(N_{s1} + P_{n1} + \frac{P_{n2}}{G_1}\right)\cdot G_1 G_2 = \left(N_{s1}\cdot F_1 + N_{s1}\cdot\frac{F_2 - 1}{G_1}\right) G_1 G_2 = \left(F_1 + \frac{F_2 - 1}{G_1}\right) N_{s1} G_1 G_2$$

This demonstrates the more general rule, $\boxed{F_T = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} \dots}$ *G is not in dB*

… the Friis formula for noise factors.[48]

The noise factor can also be considered to be the ratio of the actual output noise power to that portion of the output power due to the Johnson noise in the source resistance.

NOISE FIGURE: The Noise Figure of an amplifier (or other device) is defined as the worsening of signal-to-noise power ratio at the output compared to that at its input, expressed in decibels. It can also be considered to be the ratio, expressed in dB, of the actual output noise to the output noise of an ideal noiseless device of the same type.

$$NF = 10 \cdot \log_{10}\left(\frac{Power\ \ SNR_{INPUT}}{Power\ \ SNR_{OUTPUT}}\right) = 10 \cdot \log_{10}\left(\frac{\text{actual output noise power}}{\text{ideal output noise power}}\right)$$

$$\boxed{NF = 10 \cdot \log_{10}(F)}$$

Noise Figure is an important figure of merit for the sensitivity of a radio receiver or an RF amplifier. A receiver with a poor noise figure, for example, may not be adequate to receive a weak signal. Even a 1 dB improvement in noise figure is a substantial improvement. In digital communications systems, the Bit Error Ratio (BER) is a much more sensitive measure of the system performance than the noise figure, since a few dB change in carrier to noise ratio (caused by a noise figure change) can change the BER by several orders of magnitude (see the appendix BER vs SNR).

An unwary designer can easily be caught out by data sheets showing curves of resistance for optimum noise figure. These may lead one to think that increasing the source resistance by adding resistance somehow makes the system less noisy, since the noise figure is seen to reduce to some optimum value. WRONG! Lower noise is *always* obtained with lower source resistance. The use of these curves is for narrow band impedance matching. A matching transformer can be used to optimise the impedance driving the amplifier and therefore to improve the noise performance. If transformer isolation is not required, an LC matching network can be used to do the impedance matching. (See ***matching networks***.)

Formulae for the overall noise factor of several stages are not in decibel form. They use ***Noise Factors*** and the gains are power gains not voltage gains.

Notice that Noise Figure can be measured over a small range of frequency, giving a spot noise figure, or can be averaged over a broad range of frequency. Noise Figure takes into account both voltage noise and current noise for a device. It is great for comparing devices when the specified source impedance is what will be used in practice. If not, then you need to get the individual current noise and voltage noise values.

Although an ideal resistive attenuator does not actually add any Johnson noise to a system, its noise figure is equal to its attenuation expressed in dB. This is because noise at this attenuated position in the circuit is larger when referred back to the input.

Spectrum analyser noise figures are never stated because they are so poor. The key term to look for is *displayed average noise level*. Given that the theoretical *available noise power* at room temperature is –174 dBm/Hz, and that a greater measurement bandwidth contributes proportionately to the total noise power, the noise figure of a spectrum analyser is calculated as:

$$\boxed{\text{Noise Figure (dB)} = \text{Displayed Average Noise Level} + 174\,\text{dBm} \ \text{-}\ 10 \cdot \log_{10}(RBW)}$$

with Noise Figure in dB, Displayed Average Noise level in dBm and RBW (resolution bandwidth) in Hz. The Displayed Average Noise Level may be given over a 1 Hz resolution bandwidth, or may be expressed at say 1 kHz RBW.

[48] H.T. Friis, 'Noise Figures of Radio Receivers', in *Proceedings of the Institute of Radio Engineers*, 32 (July 1944), pp. 419-422 + correction on page 729.

A displayed average noise level of –117 dBm with a 1 kHz RBW means a noise figure of 27 dB.[†] Thus if a good wideband preamp can be used in front of the spectrum analyser, it will certainly improve the overall noise figure, albeit at the expense of worse amplitude flatness, worse harmonic distortion, worse intermodulation distortion, and so forth.

NOISE TEMPERATURE: Another way of stating the noise increase caused by a device is to quote its *noise temperature*. Continuing from the definition of ***noise factor***, all the input-referred output noise is attributed to an elevated temperature in the source resistance.

$$\boxed{F = 1 + \frac{P_{DEVICE,\ IN}}{P_{NOISE,\ IN}} = 1 + \frac{kT_{NOISE}\Delta f}{kT_{REF}\Delta f} = 1 + \frac{T_{NOISE}}{T_{REF}}}$$

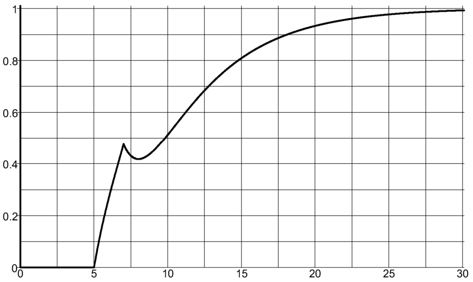
Unless otherwise specified, the reference temperature is taken as 290°K; absolute temperature units being used throughout.

Converting between ***noise figure*** and noise temperature is only slightly more complicated.

$$\boxed{NF = 10 \cdot \log_{10}\left(1 + \frac{T_{NOISE}}{T_{REF}}\right)} \text{ and } \boxed{T_{NOISE} = T_{REF} \times \left(-1 + 10^{NF/10}\right)}$$

Noise temperature is of particular interest in radio astronomy. A (lossless) antenna has a radiation resistance which is not related to the temperature of the antenna body, it is related to the source of incoming radiation. A microwave horn antenna pointing at relatively “empty” areas of space achieves a noise temperature of 3°K in the range of approximately 1 GHz to 30 GHz. Below 1 GHz the noise temperature rises by approximately two decades per decade of frequency. Above 30 GHz the noise temperature rises by one decade per decade of frequency, although atmospheric noise can be dominant.[49]

NON-MONOTONIC: A graph that is *monotonic* keeps going in the same direction, either up or down. Its rate of change in that direction may alter, but it doesn’t reverse during the measurement interval. Consider the response of a system to a step input. If the system is a simple RC time constant, the response to the step will be a smoothly rising edge. The rate of change slows, but it never reverses. The dV/dt may become 0, but it never goes negative (on the rising edge). An under-damped second order system, on the other hand, overshoots and then comes back to a steady level. The rate of change has reversed polarity. The response is *non-monotonic*.



This response is non-monotonic. If the fold-back occurs at a logic threshold then an extra transition may be detected, causing possible mis-operation of the circuit. If a ***Schmitt trigger*** is used, the hysteresis has to be larger than the fold-back in order to detect this waveform as a single transition

For a DAC wired so that increased codes give larger outputs, it is possible that at some points in the range an increase of code value will cause the output to decrease. Such a DAC is *non-monotonic*.

Certain types of DC calibrator, particularly ones with switched-resistor outputs, can give non-monotonic outputs. This is most often seen when changing the higher significant digits, such as from 8.99999 V to 9.00000 V.

NOMOGRAM: Before the age of programmable calculators and computer-based engineering software packages, the calculation of even simple formulas could be inconvenient. Charts and tables therefore gave a considerable saving of time, albeit at the expense of accuracy. A nomogram, also

[†] The best specified value for an Advantest R3172 26 GHz spectrum analyser. This drops to –106 dBm at 26 GHz (NF= 38 dB)

[49] J.D. Kraus, and D.A. Fleisch, 'Fig 5-59, Sky Noise Temperature from Radio to X-Rays', in *Electromagnetics with Applications*, 5th edn (Singapore: WCB / McGraw-Hill, 1999), pp. 336.

known as an *alignment chart*, was one such tool. A typical nomogram might consist of three parallel vertical scales; these scales not necessarily being linear. Using a ruler to produce a straight line between one point on one scale and one point on another scale, the resulting value could be read off of the third scale. Such a scale is easy to use, but not simple to devise. The method was published in detail if you should need it.[50]

NORTON EQUIVALENT: The Thévenin equivalent of a network is a voltage source in series with an impedance. The Norton equivalent [51] is a current source in parallel with this same impedance. It originated in 1926, some 43 years after Thévenin's paper. As with the Thévenin equivalent, the Norton equivalent is only suitable for evaluating what happens at the *load*. The equivalent circuit does not simulate the source in terms of internal power dissipation. The Norton equivalent only applies to linear systems, or systems that are to be considered as linear over a limited range of signal.

NYQUIST DIAGRAM: The Nyquist diagram is a polar plot of the ***loop-gain*** of an amplifier or control system. It shows ***gain margin*** and ***phase margin*** in terms of the approach of the curve to the critical $1\angle 180°$ point. This diagram is not very popular, not least of which is because polar plots are less common than Cartesian {rectangular} plots. A ***Bode plot*** conveys the same information and is easier to construct.

OCTAVE … is a frequency change of ×2. The name stems from music terminology and the Latin word *octavus* meaning eighth, as there are eight full notes in an octave. If a voltage transfer function is proportional to frequency then it will increase at a rate of 6 dB/octave. $20 \cdot \log_{10}(2) = 6.0206$. More usually, there is an AC transfer function which is inversely proportional (*asymptotically*) to some integer power of frequency. In this case the function would drop at a rate of an integer multiple of 6 dB/octave. The slope –6 dB/octave could also be expressed as –20 dB/decade.

OCXO: **O**ven **C**ontrolled (**X**)Crystal **O**scillator. There are many variants such as voltage controlled crystal oscillator (VCXO). The X is a 'cross' which sounds a bit like crys-.

OPEN-LOOP GAIN: Open-loop means "without feedback". An amplifier has a gain from its input terminals to its output terminal(s). If you apply feedback then the gain from output to input of the overall system is changed. The overall output-to-input ratio is known as the *closed-loop gain*.

To give an idea of the stability of the closed-loop response, it is not sufficient to look at the open-loop gain only; one looks at the product of the open-loop gain and the feedback factor on a ***Nyquist plot*** or a ***Bode plot***. This product of open-loop gain and feedback factor is called the ***loop-gain***.

ORTHOGONAL: The word 'orthogonal' is derived from the Latin word *orthogonius*, meaning right-angled. In this sense it is synonymous with *perpendicular*. Orthogonal vectors have a scalar product (dot product) of zero, since any 'distance travelled' in one vector direction does not increase the distance in the other vector direction.

The definition has been expanded to include function pairs which are *uncorrelated* over some specified interval. Thus the integral of the product of these functions over this specified interval is zero. The most familiar orthogonal functions are harmonically related sines and cosines. For example, $\int_0^T \sin(\omega t) \cdot \sin(n\omega t) \cdot dt = 0$, where *T* is the period of the fundamental and *n* is an integer greater than 1.

Many orthogonal series are used in filter theory and elsewhere. Examples include ***Chebyschev polynomials*** {also spelt Tschebyscheff}, Legendre polynomials, Hermite polynomials, Jacobi polynomials, Laguerre polynomials, and Gegenbauer polynomials.[52] The orthogonality relationships for these each include their own specific extra function term within the integral. For example, with

[50] R.O. Kapp, 'Nomograms in Electrical Engineering', in *Journal of the IEE*, LXXVIII (1936), pp. 567-576.

[51] E.L. Norton, 'Design of Finite Networks for Uniform Frequency Characteristic- Case 33066', *MM-1680* (Bell Labs, Nov 1926).

[52] W. Magnus, and F. Oberhettinger, 'Chapter V: Orthogonal Polynomials' in *Formulas and Theorems for the Functions of Mathematical Physics* (New York: Chelsea Publishing, 1954), pp. 78-86.

Chebyschev polynomials of the first kind, the orthogonality relation is:

$$\int_{-1}^{+1} \frac{T_m(x)\cdot T_n(x)}{\sqrt{1-x^2}}\cdot dx = 0 \ , \ \text{for m} \neq \text{n}$$

OUGHT-TO ENGINEERING: Engineers can have very arrogant ideas about what "ought to be". This is a very bad habit to get into. Let's suppose you are a worthy professional of long experience. You know that a certain piece of equipment *ought to* behave in a certain way; only an idiot would do it any other way. Well if it isn't in the spec, then you shouldn't expect it to be that way!

I have heard a customer complain that the bandwidth rolloff of their scope in the region beyond 1/10th of its bandwidth was not a monotonic single-pole response. Well who would think that it was single-pole, let alone monotonic! (Insider information from a former scope designer, me!) In reality the frequency response of a scope is never single-pole and generally becomes less likely to be monotonic as the specified bandwidth of the scope increases beyond a few hundred megahertz.

Rely on published or tested specs and not how you think something ought to behave.

OUTLIER ... is a data point which falls outside the expected range for the measurement in question. It is presumed to be a spurious data point due to some non-repetitive external interfering source, a machine being switched on for example. If there are more than a few outliers in the acquired data, the measurement system should be investigated to reduce the interference. Outliers are often discarded in order to improve measurement accuracy.

There is a good mathematical basis for asserting that outliers should be discarded. The 'noise' on a measured value might be assumed to have a Gaussian distribution. Suppose *N* nominally equal voltage readings, V_k, are taken.

The sample mean is, $\overline{\mu} = \frac{1}{N}\cdot\sum_{k=1}^{N} V_k$. The sample variance, $\sigma^2 = \frac{1}{N-1}\cdot\sum_{k=1}^{N}\left(V_k - \overline{\mu}\right)^2$,

from which the sample standard deviation, σ, is obtained. Reject individual points:

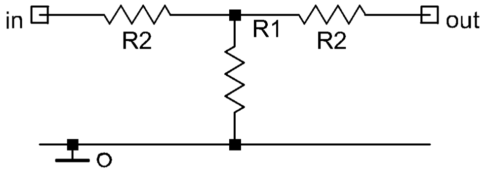
$V_k > \overline{\mu} + 6\sigma$ or $V_k < \overline{\mu} - 6\sigma$, on the basis that an individual occurrence of such a value is somewhat unlikely (1 in 500 million). The sample mean and standard deviation would be changed by removing these outliers, and that is the whole point of the exercise. You want an accurate set of data, not data skewed by some spurious event such as a power line transient. If the noise on the readings is systematic, rather than Gaussian, the peaks should be lower than those calculated, making the assumption of a Gaussian distribution a safe option. The selection of ±6σ limits is arbitrary.

PAD: In RF and microwave environments, a 50 Ω attenuator is often called a *pad*. An ideal 50 Ω 20 dB pad gives 20 dB attenuation (10× attenuation of voltage) provided that it is supplied from a 50 Ω source and is driving into a 50 Ω load.

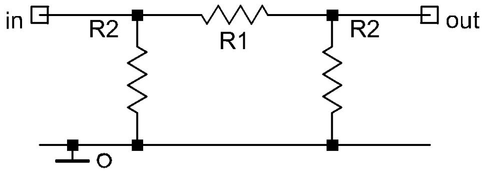
The act of using a pad improves the Z_0-matching. If an input has a return loss of 10 dB, for example, adding a (perfect) 5 dB pad will improve the return loss to 20 dB. The reflected signal goes through the pad twice; once on the way in and once on the way back out.

A correctly designed RF pad should be symmetrical, matching the characteristic impedance of the system in both directions.

Using *A* as the voltage attenuation, where $A = \dfrac{V_{IN}}{V_{OUT}},\ A > 1:$



The T-pad: $R_1 = Z_0 \cdot \dfrac{2A}{A^2 - 1}$ $R_2 = Z_0 \cdot \dfrac{A-1}{A+1}$



The Π-pad: $R_1 = Z_0 \cdot \dfrac{A^2 - 1}{2A}$ $R_2 = Z_0 \cdot \dfrac{A+1}{A-1}$

A 3 dB pad ideally gives 3 dB insertion loss and 6 dB improvement in return loss. If it is necessary to get less insertion loss, but more improvement in the return loss then an ***isolator*** should be used. Figures of 1 dB insertion loss and 18 dB return loss are possible using a suitable isolator, but only over an octave of bandwidth.

To pad has an additional definition. It means to add more of something in order to increase the total. A typical example would be in an attenuator with two switched paths. If one path has a lower input capacitance then it may be desirable to *pad* the capacitance up to the same value as the other path. Fundamentally the process is one of adding something to improve the matching or balance of a circuit.

The word *pad* is also used in software when extra characters are used to increase a data stream to an appropriate size.

PARASITICS ... are undesirable features of components or subsystems which tend not to be reducible by the techniques of guarding, shielding and component positioning.

Examples of parasitic effects are:

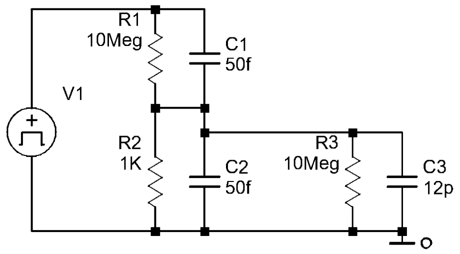
☹ lead or track inductance in a resistor
☹ lead or body inductance in a capacitor
☹ lead resistance in a capacitor
☹ inter-terminal capacitance in an IC or a transistor
☹ coil-to-contact capacitance in a reed-relay

The common factor is that an undesirable ‘feature’ exists within the component and it is necessary to redesign the component to reduce the effect. PCB effects are in a grey area between *stray* and parasitic effects.

Wire-ended components typically have higher parasitics than surface mount parts. The lower parasitics, and the smaller size, are two of the reasons why modern designs for RF circuitry above a few tens of megahertz have to be done using surface mount parts. Even surface mount parts have series inductance and shunt capacitance however. Typical values for surface mount resistors are:[53]

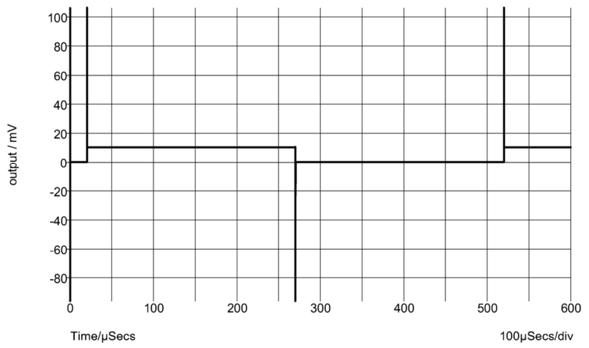
SIZE	SERIES-L	SHUNT-C
1206	2 nH	0.05 pF
0805	1 nH	0.09 pF
0603	0.4 nH	0.05 pF

[53] 'Introduction, Chip Resistors', *Discrete Ceramics* (Philips Components, May 2000).



These parasitics can have effects at surprisingly low frequencies. This is a simple attenuator where C1 is the parasitic capacitance of R1, and C2 is the parasitic capacitance of R2. The attenuator is monitored by a 10:1 scope probe represented by R3//C3.

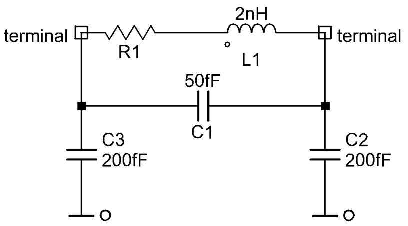
The simulation of this circuit with a 100 V 2 kHz 50 ns risetime square wave input is startling at first sight.



This result is not what you might have expected from the attenuator. Since all components have such parasitic elements, it is usual when designing attenuators to include capacitors across the resistive elements in order to define what happens even at moderate frequencies.

The presence or absence of a ground plane, a shielding box, or tracks nearby, will give capacitance to other signal tracks or to ground. Capacitance to other signal tracks at best gives you a path to signal ground, and at worst gives you *cross-talk*. This extra capacitance can easily exceed the parasitic 50 fF capacitance by an order of magnitude (ie 0.5 pF to ground).

Components need to be held in place and wired together. This requires a strong, solid material, which must necessarily have a dielectric constant greater than unity. Typically this material would be PCB laminate or some type or ceramic. Whilst ceramic has excellent heat spreading qualities, it has a dielectric constant in the region of 6 to 10. Ordinary FR4 PCB material has a dielectric constant in the region of 4 to 6. It is possible to obtain expensive PCB materials for operation above a few hundred megahertz, and having a dielectric constant around 2, but they are expensive, and are not nearly as strong as standard FR4.



This is a realistic model for a 1206 surface mount resistor soldered to a PCB. C2 and C3 are representative values for the stray capacitances of the resistors' terminals. Clearly the values will change according to the proximity of the ground plane and/or other tracks.

Historically, putting several lower value resistors in series produced one larger value resistor with a reduced time-constant of the shunt capacitance. This scheme fails spectacularly if the stray capacitance to ground has any significant value and is definitely not workable for surface mount components. The resultant distributed time-constant ruins the attenuation characteristic of the overall network. The best practice is therefore to use as few resistors as possible in an attenuator design. This means pushing resistors right up to their power and voltage limits.

One can also reduce package parasitics by removing the package! The bare silicon die can be stuck directly to the PCB and wire-bonded to it. This is a specialised and expensive process. There is a trend towards reducing packaging parasitics by taking well passivated silicon chips {dice} and flipping them over before directly soldering them to the PCB.

PEDESTAL OFFSET: In sampling systems such as ADCs, CCDs, and sample & holds, the sampling system always captures some of its own clock. This is not a problem if the amount captured stays constant. The problem comes when there is a shift of the delay between the sampling clock and the feedthrough signal, perhaps as a result of a changing ambient temperature. The result is a drifting DC offset. The name comes from the fact that the sampled data is standing on a DC pedestal {a base, support or foundation}. The pedestal may also change with the rate at which samples are taken. This then gives a sampling rate induced DC offset error.

PELTIER EFFECT: Discovered in 1834 by the French physicist J. Peltier, this is the reverse of the

Seebeck Effect. If a current is passed through a series connected pair of junctions of dissimilar metals, one junction will be heated and one will be cooled. If many such junctions are used to make a *thermopile*, it is possible to make a Peltier-effect heat pump (Thermo Electric Cooler, TEC). These devices are most efficient when made in semi-conducting materials.

A heavy load on a thermocouple will therefore change the measured point by a slight amount. The current tries to equalise the temperatures by taking the heat from the hotter junction to the colder junction. Such an error is not important in thermocouple measurement applications because the current is minimal (<<1μA).

PER-UNIT … is just the ratio of a change to the original amount. It is a more basic form than percent. If a resistor changes from 2.03K to 2.27K, it has changed by 0.24K in 2.03K. This is a per-unit shift of 0.118; a percentage shift of 11.8%.

Per-unit is a dimensionless form and can be used directly in many applications. It is therefore often preferable to percent. Multiply the per-unit value by 100 to get the percent value.

PHASE NOISE: A pure oscillator would produce only the required frequency and nothing else. No real oscillator can achieve this due to noise in the system. This noise appears as small variations in the instantaneous output frequency and amplitude. If the oscillator output is viewed on a spectrum analyser, the amplitude will be seen to fall away rapidly on both sides of the oscillation frequency. However, some of this finite roll-off may be due to the characteristics of the spectrum analyser's ***intermediate frequency*** (IF) filters.

If the spectrum analyser's local oscillator is better than the oscillator under test, the phase noise can be read off the spectrum analyser screen as follows: Phase noise is conventionally given in units of dBc/Hz. This is dB relative to the carrier [largest signal] in a 1 Hz bandwidth. Direct readouts from spectrum analysers are often in dBm. In this case read the amplitude of the carrier and subtract the reading at the chosen measurement frequency. This will give a reading in dBc. Alternatively the spectrum analyser may have cursor readouts directly in dBc. The *offset frequency* is the frequency difference between the measurement frequency and the carrier frequency. Now read the *resolution bandwidth* (RBW) figure on the spectrum analyser. Suppose you get –65 dBc at 10 kHz offset with a resolution bandwidth of 1 kHz. Dividing the power by the resolution bandwidth is equivalent to subtracting $10 \cdot \log_{10}(\mathrm{RBW})$ from the dBc figure. Hence the phase noise is found by subtracting $10 \cdot \log_{10}(1000)$ from –65 dBc, giving –95 dBc/Hz @ 10 kHz offset.

Notice that a spectrum analyser with a sharper (more selective) IF filter allows you to get a lower phase noise measurement for any given setup of resolution bandwidth. This selectivity may be quoted in terms of the 60 dB / 3 dB ***shape factor***; 15 being ok, 11 being good, and 5 being excellent.

If the display on the spectrum analyser is a smooth curve, coming down symmetrically on both sides of the carrier, you are just looking at IF filter characteristic of the spectrum analyser. Reduce the resolution bandwidth until the response curve actually looks *noisy*; what remains is the actual system noise, not the IF filter selectivity curve.

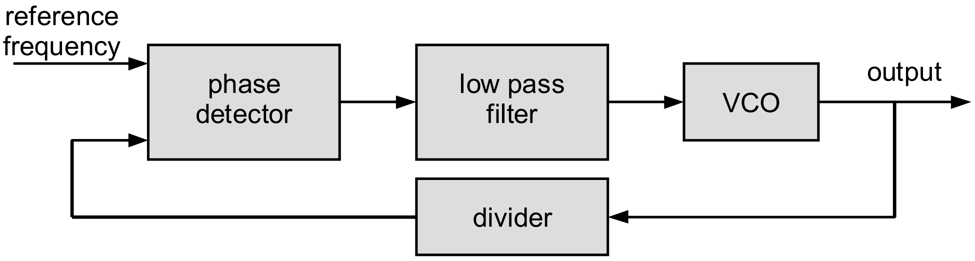
PHASE MARGIN: In control theory, and in amplifier feedback systems in general, a system becomes unstable if the ***loop-gain*** magnitude reaches unity with 180° phase shift. The phase margin is the amount by which the loop phase shift is less than 180° when the loop-gain magnitude is 1. In order to get a good pulse response from a system it is necessary to have both a good ***gain margin*** and a good phase margin.

On a system where the loop-gain approaches 0 dB on a shallow curve, the gain margin is far more important than the phase margin. In any case a phase margin of at least 45° is desirable, as is a gain margin of at least 10 dB. These values are only a starting point for a design; the closed-loop frequency response and pulse response tell you more about the system stability than these theoretical measures. The important thing to know is that a peaking frequency response is due to either inadequate gain margin or inadequate phase margin, or both.

PIM: **P**assive **I**ntermodulation **D**istortion. ***Intermodulation Distortion*** is where the simultaneous presence of two or more signals causes the creation of additional frequencies due to the non-linearity in the system. This is a well known problem in amplifiers, but is very much less well known in attenuators, connectors and cables. It is also much more difficult to measure in cables and connectors because the distortion levels are up to 100 dB lower than in amplifiers.

Passive Intermodulation Distortion is simply intermodulation distortion in these passive devices. As you can imagine, the level of PIM in connectors and cables is pretty low because the power loss and non-linearities are very low. The PIM in cables and connectors is, however, of great importance in RF transmitters, where there can be hundreds of watts of power passing through the cables. Testing cables and connectors up to and beyond −160 dBc requires dedicated test fixtures and considerable expertise.[54]

PLL: Phase **L**ocked **L**oop. The three essential elements of a phase-locked loop are a phase comparator, a filter, and a voltage controlled oscillator (***VCO***). When the loop is 'locked', both inputs to the phase detector will be at the same frequency. The filtered output of the phase detector will be at a fixed DC level.



In this PLL block diagram, an (optional) frequency divider has been added to make the output frequency higher than the reference frequency by the frequency division ratio of the divider. This multiplication of the output frequency is useful because high stability reference frequencies are often derived from crystal oscillators below 20 MHz. High accuracy outputs at arbitrarily large frequencies can therefore be achieved by this multiplication technique.

There are two distinct types of application for phase-locked loops. One is a static use, where the reference frequency is fixed. The other is the dynamic case, where the phase-locked loop is tracking the incoming signal in order to demodulate a phase or frequency modulated signal.

For the static case, the time taken for the loop to settle is often not critical. The primary considerations would be: does the loop lock at all, and is the ***phase noise*** acceptable? To attain minimal phase noise the frequency span of the VCO should be as small as practicable. In the case of a >20 MHz LC oscillator locked to a crystal, the input frequency is known to ±100 ppm worst case. Allowing ±15% tolerance on the inductor and capacitor gives at least a ±15% tolerance on the frequency, not allowing for the variation caused by the amplifier. If the VCO is capable of taking out this ±15% tolerance it will generally be several times noisier than necessary.

Look at the system this way: The control voltage span of say 5 V is set to give an adjustment range of say ±20% of the output frequency. 10 mV of noise on this control voltage will therefore try to change the frequency by 0.08%. The loop will then react to remove this error and the result will be phase noise. If the adjustment range is reduced by a factor of 4 to say ±5%, the phase noise produced by this noise mechanism will be proportionately reduced. This reduction is achieved by tighter initial component selection or by in-circuit adjustment. There is clearly a trade-off between the phase noise at the output and the additional cost of adjustments.

It is a mistake to think that "the loop" will remove power supply noise from the oscillator output. It is true that the loop will adjust the VCO to try to stabilise the output frequency. Consequently a moving signal on the VCO input, given a stable reference frequency, is indicative that the loop is working. However, if it is having to work, the phase noise resulting from the corrective process may be significant in your application. Reduce the movement of the VCO input by better power supply decoupling and better shielding of the oscillator.

Measuring the VCO input on a scope, possibly using an active probe to minimise probe loading effects, is a quick, powerful and quantitative way of seeing how much improvement your decoupling or shielding is having on the PLL. However, a word of warning, never try to filter the VCO input signal

[54] P Ling, 'On the Same Wavelength', in *New Electronics*, Feb 2000, pp. 57-58.

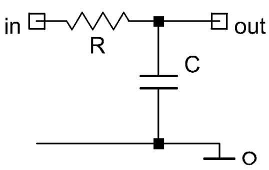
to remove this “noise”. The signal on the VCO input is required to stabilise the loop. If you applied a fixed DC level to the VCO input, the frequency output would be considerably noisier.

To debug a PLL, disconnect the phase detector output from the VCO input. Drive the VCO input from a pot to make sure the VCO has enough range. At the same time monitor the filtered phase detector output. The phase detector output should be the difference frequency between the reference frequency and the VCO. This beat frequency must not reduce in amplitude over the range of possible beat frequencies; reduction in amplitude implies phase shift and this could cause loop instability. If the loop does not work when connected up, suspect that the phase of the feedback signal is reversed.

POLARISATION … means that there is direction involved in a component or a situation. An electrolytic capacitor is *polarised* in the sense that the + end has to be held more positive than the other end. Connectors may be *polarised* meaning that they can only be inserted one way round.

Electromagnetic radiation can be polarised meaning that the electric field is oriented in a specific way. For plane polarised electromagnetic radiation, it is the plane of the electric field which defines the plane of polarisation. The magnetic field is perpendicular to the plane of polarisation. The electric field distant from a dipole is parallel with the electrodes. Thus a horizontal dipole transmits or receives horizontally polarised radiation.

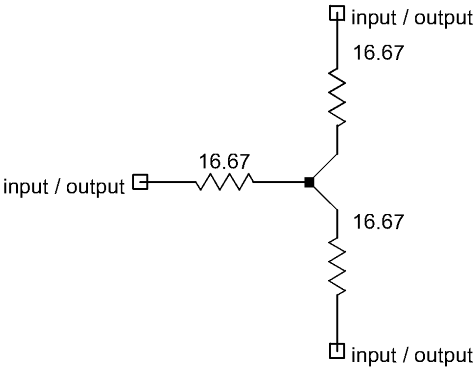
POLE: This is a single-*pole* filter. The term pole comes from the plot of the transfer function,



$T = \frac{1}{1+sCR}$, against the complex variable *s*. When *s* is set to $\frac{-1}{CR}$ the value of *T* becomes infinite. The magnitude of *T* plotted against the complex variable *s* gives a two-dimensional surface. This surface can be visualised as canvas, with a tent pole pushing the canvas up to infinity when $s = -\frac{1}{CR} + j \cdot 0$. The term ‘pole’ is just a contraction of this tent pole.

The matching term is *zero* for those points where the value of *T* becomes zero. This would be due to a zero factor in the numerator {top part} of the transfer function.

POWER DIVIDER:



This is a fully Z_0-matched device with three identical ports. It is usual to apply a signal to one port and have it halved in voltage terms when it appears at the other two ports, thereby giving a 6 dB loss to each of the ports. For a 50 Ω system, each resistor has a value of exactly $\frac{50}{3}\Omega$. In order to avoid confusion with a ***power splitter***, it is better to call this one a 3-resistor power divider.

POWER FACTOR: There is *real power* {average power; mean power; actual power; true power} and there is *apparent power*. The apparent power is the product of the RMS current and the RMS voltage. (Anyone found talking seriously about “RMS power” should re-study first year EE textbooks.)

For sinusoidal waveforms: $\boxed{\text{Real Power} = V_{RMS} \cdot I_{RMS} \cdot \cos(\phi)}$, where ϕ is the phase angle.

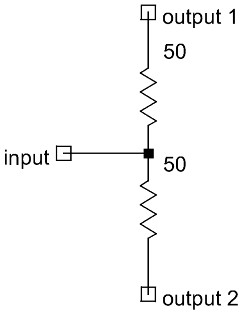
In general, neither the supply voltage nor the load current need be sinusoidal, so the definition of power factor as the cosine of the phase angle between the voltage and current is often inapplicable.

$$\boxed{\text{Power Factor} = \frac{\text{Real Power}}{\text{Apparent Power}} = \frac{\frac{1}{T}\int_0^T v(t) \cdot i(t) \cdot dt}{V_{RMS} \cdot I_{RMS}}}$$

where *T* is the repetition period of the waveform.

Harmonic currents drawn from a harmonically pure supply voltage cannot supply power to the load; all they do is dissipate power in the distribution network. Some harmonics are much more intrusive to a 3-phase power distribution network than others because the currents add in-phase. These are the ***triple-n harmonics***, also written as *triplen*; harmonics which are odd multiples of the third harmonic [3rd, 9th, 15th, 21st ...].[55]

POWER SPLITTER: The circuit shows a 50 Ω power splitter. The signal is applied to the input and splits equally to the two outputs if they are properly Z_0-matched. Notice that this splitter is only Z_0-matched in the forward direction. The reverse Z_0-matching is best when the source connected to the input has zero output impedance.



If one of the outputs is a Z_0-matched sensing device, such as a power meter or a spectrum analyser, then the splitter can be used as a sensing device in a feedback loop to deliver a signal to the other output with a low effective output mismatch {high return loss; low VSWR}. Such a configuration is useful because amplitude levelled sources {signal generators} with low output VSWR are not readily available even at modest RF frequencies, let alone microwave frequencies.

In order to avoid confusion with a ***power divider***, it is better to call this device a “2-resistor power splitter”. Like the power divider, the power splitter also gives 6 dB insertion loss and is therefore only 50% efficient. Matched lossless power splitting requires matching transformers or their equivalent, for example a ***Wilkinson divider***.

POWER SUPPLY REJECTION: The extent to which power supply voltage does not affect a circuit. The finite power supply rejection ratio (***PSRR***) of an amplifier means that some of the noise on the power rail[s] will find its way onto the output of the amplifier. Power supply rejection ratio normally gets worse with frequency, often at a rate of 20 dB/decade. The power supply rejection ratio is always referred to the input of an amplifier for spec purposes.

PPB: **P**arts **P**er **B**illion. This is not a safe term to use without further clarification and should therefore be avoided. In the UK and Germany a billion is one million millions (10^{12}). In the US, Canada and France it is one thousand millions (10^9). This means that the term ppb can be mis-interpreted. If you read articles with ppb in them, the best guess would be that it means parts in 10^9 unless there is other evidence to the contrary.

PPM: **P**arts **P**er **M**illion. ppm=1,000,000 × ***per-unit*** ; ppm=10,000 × percent. Where accurate and/or small values are involved, it is more convenient to work in ppm than in percent. For example a resistor shift from 3.0166K to 3.0168K is 0.0066% or 66ppm. A simple rule for voltage is that 1 ppm is 1 μV/V.

Formal scientific notation uses terms like 1 part in 10^7 rather than 0.1ppm. One reason for this is that 100 ppm does not automatically mean that the last two zeros are *significant*. If you saw 124 ppm you could be confident that you were being given the number with 3-digit precision. When the number is 100 ppm, you cannot be sure if this means that there is only one digit of precision. This difference could be highlighted implicitly by saying either 1 part in 10^4 or 100 parts in 10^6.

ppm: is also used for printers and photocopiers where it means **p**ages **p**er **m**inute

PRE-ARCING TIME: For a fuse, pre-arcing is the time from the application of an overload current to the instant when an arc starts. The total time taken to *clear* the fuse (break the circuit) is the sum of the pre-arcing time and the arcing time. Fuses are very effective isolating devices, but the time taken to ‘blow’ the fuse can get excessive if the fault current is only a few times the rated current.

PROXIMITY EFFECT: The ***skin effect*** means that for all frequencies, current density is higher at the outer surface of an isolated conductor. At power frequencies, 50 Hz / 60 Hz, the skin depth of

[55] J. Shepherd, A.H. Morton, and L.F. Spence, '5.10: Harmonics in Three-Phase Systems', in *Higher Electrical Engineering*, 2nd edn (ELBS & Pitman, 1970; repr., 1975), pp. 153-155.

around 9 mm (in copper) may be larger than the wire radius; the variation of current density with distance from the surface of the wire will not then be particularly important. However, it should be clear that increasing a power cable radius above a few centimetres will not have the desired effect of further reducing the volt drop in the cable. For a given cross-sectional area of conductor, the most efficient cross-sectional shape is a hollow circular tube.

If current carrying conductors are placed near to each other, the currents will be attracted or repelled according to the relative phases of the currents. At DC the rule is that 'like currents attract', the opposite effect to that observed for individual charges. The redistribution of current that results, restricts the current to a smaller area of conductor and therefore increases the effective resistance of the conductor. This redistribution of current is known as the *proximity effect*. As an example, if tubular conductors are placed almost touching each other, the power loss can be twice that achieved by leaving a gap between the conductors equal to their diameters.

In an isolated rectangular conductor the current distribution gets even more 'bunched-up'. The current density is pushed out to the corner surfaces of the conductor, thereby increasing the effective resistance to a greater extent than the simple skin effect model would predict.[56]

This proximity effect is the reason why inductors have a higher Q when the turns in the winding use a smaller gauge of wire than is necessary to fill the space. If the turns are too close to one another, the loss due to the proximity effect is greater than the gain due to the increased wire diameter. The gaps between wires can be made equal to or greater than the wire diameter in order to improve the Q.

Switched-mode transformer efficiency is also greatly affected by proximity effect losses. Proximity effect losses can be reduced by minimising the number of winding layers and by interleaving the primary and secondary windings. However, the optimisation of wire gauge, wire separation, number of layers, ratio of copper volume to iron volume &c is more to do with skilled insight than plugging numbers into formulae.

Niobium is a superconductor, but copper is not. However, if a copper conductor is in intimate contact with a Niobium superconductor, the Niobium can induce superconductivity into the copper. Thus when dealing with superconductors, the term 'proximity effect' can relate to this induced superconductivity effect.

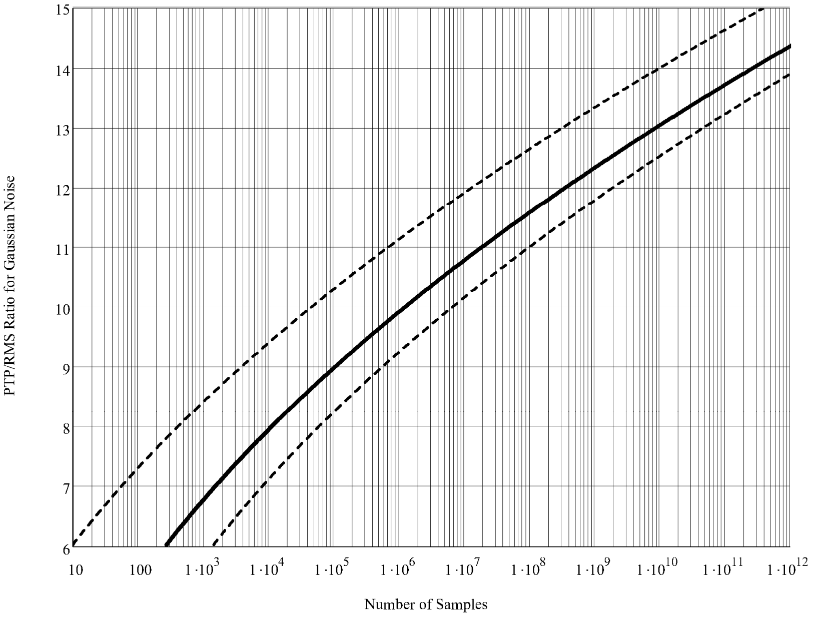
PSRR: Power **S**upply **R**ejection **R**atio is the change in input offset voltage with supply voltage for an amplifier or other signal conditioning device. More practically it can be considered to be the input referred noise caused by power supply noise. Often one power rail has a lower PSRR than the other, sometimes by as much as 20 dB.

$PSRR = \frac{\Delta\text{supply ripple}}{\Delta\text{input referred noise}}$ or in dB form $\boxed{PSRR\,(\text{dB}) = 20 \cdot \log_{10}\left(\frac{\Delta\text{supply ripple}}{\Delta\text{input referred noise}}\right)}$

PTAT: Proportional **T**o **A**bsolute **T**emperature. The transconductance of a silicon bipolar transistor is $g_m = \frac{I_C}{V_T} = \frac{I_C}{kT/e}$. As the transistor warms up, the transconductance decreases. Since temperature-stable gain of an amplifier is highly desirable, it is useful to keep the I_C/T ratio constant. This can be done by biasing the stage from a current source which increases *proportionately to absolute temperature* {temperature measured in °K}. Such a current source is called a PTAT current source, or just PTAT.

PTP: Peak-**t**o-**p**eak. A typical scope measurement used to specify the amplitude of a waveform rather than using RMS. Also written as ptp and p-p. PTP readings are very susceptible to noise and the peak-to-peak reading increases (slowly) with the interval of observation. For true Gaussian noise that is sampled, the mean value of the peak-to-peak reading can be estimated from the graph overleaf:

[56] A.E. Kennelly, and H.A. Affel, 'Skin-Effect Resistance Measurements of Conductors at Radio-Frequencies up to 100,000 Cycles Per Second.', in *Proceedings of the IRE*, 4 (1916), pp. 523-580.



The mean ratio of PTP/RMS is given by the heavy middle line. The lower line is a 2.5% probability and the upper line is a 97.5% probability. Thus there is a 95% chance that the PTP reading will be found between the two outer lines.[57]

If there is 1 μV RMS of random noise in the bandwidth of a sampling system with adequate resolution, and 10,000,000 samples are taken, the ptp reading should be between 10 μV and 12 μV approx.

You will often come across the use of 6σ, where σ is the standard deviation. Standard deviation and RMS value are essentially the same for this application. You can see from the curves that 6σ is only representative of the ptp spread when a relatively small number of samples are taken.

Some people seem to think that if you sample for a very long time the noise will be infinite! This is arguably true in some obscure mathematical sense, but as the curves show, the increase in ptp value is very slow. Values above 20σ are not very real. Furthermore, in real life noise is not quite as "Gaussian" as this curve predicts. In practice, real noise tends to err on the low side of the above curves (lower PTP value for a given RMS value).

QUARTER-WAVE TRANSFORMER … is a *transmission line transformer*, not a wound coil transformer. The idea is to convert a low impedance into a high impedance, or vice-versa, although this transformation only occurs over a limited frequency range.

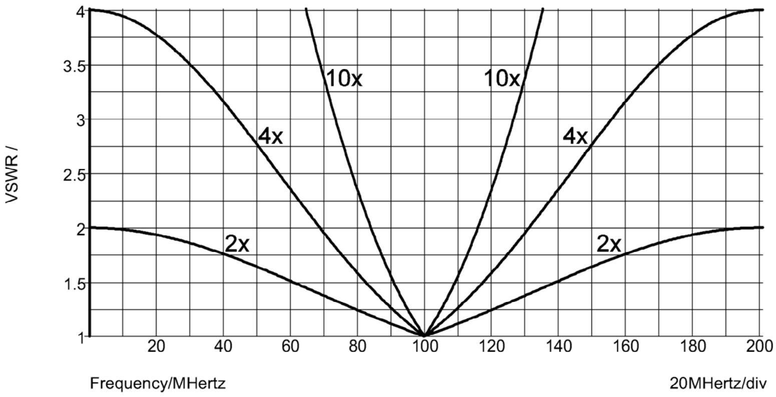
If a lossless transmission line is loaded by a mis-matched resistor, $R_L \neq Z_0$, the VSWR on the line will be $\frac{R_L}{Z_0}$ for $R_L > Z_0$ and $\frac{Z_0}{R_L}$ for $R_L < Z_0$. For a 50 Ω line and a load VSWR of 2, the line will 'look like' 100 Ω at one distance from the load, and 25 Ω at another distance. Thus a low value resistance can be transformed to a high value resistance by a transmission line of length $\lambda/4$, a *quarter-wave line*. The same transmission line will also transform a high value resistance into a low value resistance. On a *Smith Chart* this is equivalent to moving 180° around the chart.

$$R_{HIGH} = VSWR \times Z_0 \quad ; \quad R_{LOW} = \frac{Z_0}{VSWR}$$

$$R_{HIGH} \times R_{LOW} = VSWR \times Z_0 \times \frac{Z_0}{VSWR} = Z_0^2 \qquad \text{Hence} \quad \boxed{Z_0 = \sqrt{R_{HIGH} \times R_{LOW}}}$$

The characteristic impedance of the transmission line is made equal to the geometric mean of the desired high and low resistance values. This technique only works over a limited band of frequencies, the band being reduced by the amount of impedance transformation required. The acceptable band is also reduced if a low input VSWR is specified.

[57] L.O. Green, 'Getting the Most from Your Scope', in *Electronics World*, 106, no. 1769 (May 2000), pp. 384-386.

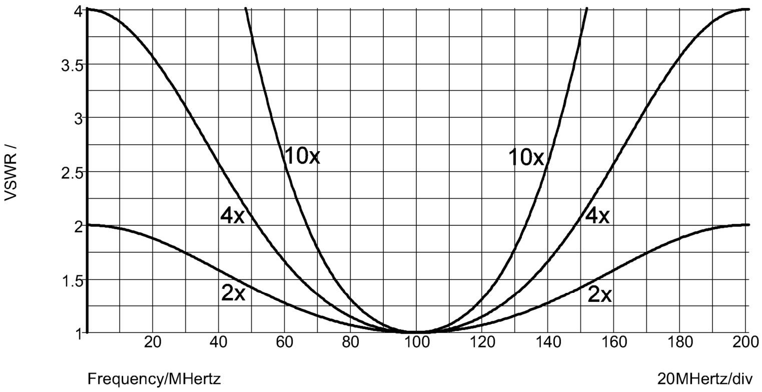


The quarter-wave transformers plotted are for operation 100 MHz; this makes them 2.5 ns long. The input VSWR curves for the quarter wave transformers are for load resistances 2×, 4× and 10× the desired input resistance. The input VSWR curves for resistive loads lower than the characteristic impedance are identical to those shown for a given VSWR load.

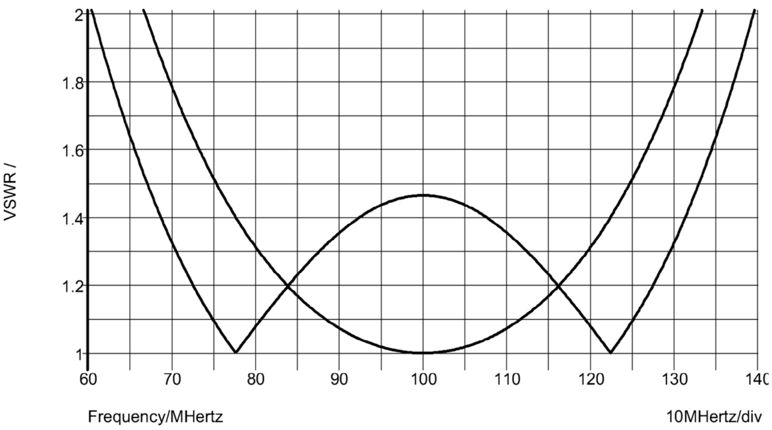
The curve for a 100 Ω load is the same as that for a 25 Ω load.

Notice that if the acceptance criterion is a VSWR of 1.5, for example, the higher VSWR loads are matched over a considerably reduced frequency range. A larger acceptable frequency range is achieved by using two quarter-wave transformers in series.

Quarter-wave transformers matching 100 Ω, 200 Ω and 500 Ω loads to a 50 Ω input:



The transformer values are: $Z_{LO} = \sqrt{\sqrt{R_{LOW}^3 \times R_{HIGH}}}$ for the line connected to the low source resistance, R_{LOW} , and $Z_{HI} = \sqrt{\sqrt{R_{LOW} \times R_{HIGH}^3}}$ connected to the load R_{HIGH} .



The biggest ratio of increase in acceptable bandwidth, compared to the single quarter- wave transformer, is achieved for the highest VSWR load matched (500 Ω).

If some non-ideal VSWR is acceptable at the centre frequency, then the usable range can be extended by adjustment of the impedances of the lines.

In this example of a 500 Ω load in a 50 Ω system, the impedance of the line near the source has been multiplied by 1.1 and the impedance of the line near the load has been divided by 1.1.

By comparison with the unmodified version, the modified version is seen to have a non-monotonically increasing VSWR, but an improved acceptable frequency range as a result.

In general, the line impedances are moved closer together by the same factor to maintain symmetry with respect to the centre frequency.

$$\boxed{Z_{LO} = x \cdot \sqrt{\sqrt{R_{LOW}^3 \times R_{HIGH}}}\;; \quad Z_{HI} = \frac{1}{x} \cdot \sqrt{\sqrt{R_{LOW} \times R_{HIGH}^3}}\;; \quad x \geq 1\;; \quad VSWR_{f_0} = x^4}$$

The increased acceptable frequency range is conceptually similar to the improvements in filter roll-off that can be achieved when ripple is allowed in the passband.

Three $\lambda/4$ transformers can achieve a slightly larger bandwidth than the correctly matched two transformer solution, with the ultimate being a line having a smoothly tapered impedance.[58]

A quarter-wave transformer can also be used as an RF constant current source over a restricted range of frequencies. If one end is fed from a voltage source having zero ohms source impedance, the voltage across a resistive load is proportional to the resistance.[59] On a lossless 50 Ω line driving a 50 Ω load the voltage output will equal the voltage input.

Thus, the current source has a value $\boxed{I_{OUT} = \frac{V_{IN}}{Z_0}}$.

A quarter-wave transformer is always better than a single *stub-tuned* circuit.

A transparent quarter-wave coating on a quasi-optical mm-wave component can be used to eliminate reflection from the component, the requirement being that the intrinsic impedance of the coating is the geometric mean of the free space impedance and the impedance of the mm-wave component.

QUASI-STATIONARY STATE: Also known as the *quasi-static state*, quasi- meaning partly or almost. Any electrical field problem can, in principle, be solved by direct application of *Maxwell's equations*. In practice, however, such solutions are only obtainable in certain geometrically simple situations. It is therefore often convenient to neglect the complexity of a complete and rigorous solution to Maxwell's equations by using a "lumped equivalent circuit".

It is usual in introductory courses to consider electrical circuits in terms of resistances, inductances and capacitances. The assumption is that the dimensions of the components are small compared to the wavelength of the applied signal. This is necessary because the velocity of propagation of the electromagnetic energy is required to have a negligible effect on the overall result.

Consider an air-wound {air spaced; formerless} inductor supplied at such a high frequency that the length of wire is comparable with the wavelength of the applied signal. In this case the phase of the current at one end of the inductor will be significantly different to that at the other end. Thus the simple idea of inductance as the ratio of linked flux divided by the current becomes inapplicable.

When either of the terms *quasi-stationary* or *quasi-static* are used, they are intended to mean that the system is running at a sufficiently low frequency that the phase of any current is almost constant throughout an individual component part. Obviously one is free to choose the amount of phase error that can be considered as tolerable, therefore there is no precise frequency at which a system is definitively quasi-static.

RADIATION RESISTANCE: RF current fed into an antenna will produce heat and RF radiation, both of which can be modelled as series resistances.

$$\boxed{\text{radiation resistance, } R_R \equiv \frac{\text{total radiated power}}{(\text{antenna input current})^2}}$$

The antenna input current is the RMS value of a pure sinusoidal signal.

[58] J. Willis, and N.K. Sinha, 'Non-Uniform Transmission Lines as Impedance Transformers', in *Proceedings of the IEE*, 103, part B (1956), pp. 166-172.

[59] E.A. Guillemin, 'Chapter 2-13: Odd and Even Quarter-Wave-Length Lines', in *Communication Networks Vol. II, The Classical Theory of Long Lines, Filters and Related Networks.* (Wiley, 1935; repr., 1953), pp. 64-65.

Write R_R = radiation resistance; R_L = loss resistance,

$$\boxed{\text{radiation efficiency}, \eta = \frac{R_R}{R_R + R_L}}$$

Mismatch at the antenna terminals is not relevant to this definition. Note that in general there would also be a reactive part to the antenna impedance. Thus $Z_A \equiv (R_R + R_L) + jX_A$

The radiation resistance of a *small* loop (of any shape) having *N* turns is $\boxed{R_R = 320\pi^4 N^2 \left(\frac{A}{\lambda^2}\right)^2}$

where *A* is the area of the loop and λ is the free-space wavelength of the radiated signal.

In order to be considered small, the length of wire in the loop has to be much shorter than a quarter wavelength long: $N \times \text{loop perimeter} < \lambda/4$

The radiation resistance of a short dipole $\left(L < \lambda/10\right)$ is: $\boxed{R_R = 20\pi^2 \left(\frac{L}{\lambda}\right)^2}$, where *L* is the tip-to-tip length of the dipole. This equation assumes a triangular current distribution in the antenna, the current falling to zero at the tips. If the current distribution is uniform due to a capacitive end-load, the radiation resistance increases by a factor of ×4.

The radiation resistance changes significantly when an antenna is brought close to conducting objects such as metal sheets or the Earth. The reason is that there is an image antenna formed in the ground plane, the radiation from which adds to or subtracts from the transmitted wavefront, depending on the height above the ground plane and the direction being tested.

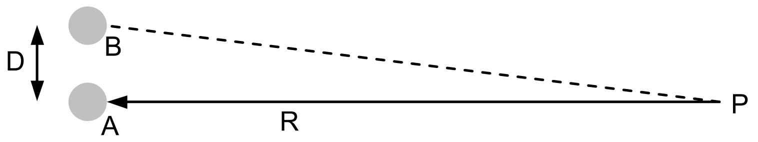
Quotation: [60]

> **"… a thin steel wire or a wet string have the same radiation resistance as a large copper bar."**

Note that the wet string has a much lower *radiation efficiency* (see above).

RAYLEIGH DISTANCE … originates from optics where the idea is to be able to *resolve* sources, meaning to separate the images such that individual sources are just perceptible.

If the small spheres represent in-phase sources of a signal frequency, the path difference between PA and PB will cause a phase shift. The short path is R, and the longer path is $\sqrt{R^2 + D^2}$. The Rayleigh distance is defined by the criterion that the path difference between the long and short routes is equal to $\lambda/4$.



Hence $\sqrt{R^2 + D^2} - R = \frac{\lambda}{4}$ giving $\sqrt{1 + \frac{D^2}{R^2}} - 1 = \frac{\lambda}{4R}$

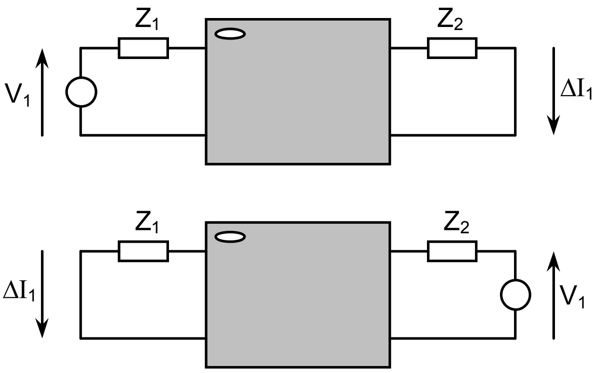
If it is specified that $\frac{D}{R} << 1$, the first order binomial expansion of the square root gives

[60] C.P. Steinmetz, 'The General Equations of the Electric Circuit III; Variation of Constants r, L, C, and g, and Its Effects.', in *Proceedings of AIEE*, 38 (Feb 1919), pp. 249-318.

$\frac{D^2}{2R^2}=\frac{\lambda}{4R}$ and finally $\boxed{R=\frac{2D^2}{\lambda}}$, where *R* is the Rayleigh distance.

The above argument assumed that the receiving device was negligibly small compared to the transmitting source. A better formula takes account of both the source and the receiver, $R\geq\frac{2(D_1+D_2)^2}{\lambda}$. It is good idea to separate the antennas by more than the formula suggests, if possible, in order to minimise the phase shift and therefore to minimise the amplitude reduction. Furthermore, if a receiving antenna is placed within the Rayleigh distance the back-scatter from the receiving antenna can mismatch the transmitting antenna, causing measurement uncertainties.

RECIPROCITY THEOREM … is a very important and wide-ranging part of linear passive network theory. If a current change ΔI_1 occurs in branch A of a network, due to an additional voltage V_1 in branch B of the network, then this same additional voltage V_1 in branch A will cause a current change ΔI_1 in branch B. Reciprocity will occur with DC or AC signals. This form of the reciprocity theorem first appeared in Maxwell's treatise,[61] although Lord Rayleigh has been credited with noting that reciprocity applies when the internal coupling is by induction.[62] Maxwell credits Professor Felici of Pisa with experimentally demonstrating reciprocal induction as early as 1859.[63]



Without loss of generality this situation can be viewed as a two-port network having externally placed impedances Z_1 and Z_2 as shown.

There is no requirement that Z_1 and Z_2 should be equal, or indeed non-zero. There is also no requirement for a galvanic connection (DC path) between input and output.

For antennas, this ordinary definition of reciprocity is not convenient because one cannot get voltage generators or ammeters with zero impedance. However if both the source and the receiver are matched to the same characteristic impedance, one does get reciprocity of the input to output voltage, current and power. Under these conditions in a radio communication link, interchanging the transmitting and receiving antennas will not change the received signal.[64]

It is essential that both antennas remain linear during such a test. In a large antenna with badly made joints, it is possible for the joints to have a high resistance during reception and a low resistance during transmission, the poor quality connection arcing over during transmission to form a low resistance path. In this case, transmission is far more effective than reception;[65] the antenna is decidedly non-linear and non-reciprocal.

An additional requirement for reciprocity with antennas is that the transmission path remains linear. Such linearity is not guaranteed if either antenna requires a balun, or if the transmission passes

[61] J.C. Maxwell, 'System of Linear Conductors', in *A Treatise on Electricity and Magnetism*, 1st edn (Oxford: Clarendon Press, 1873), pp. 335 (para 281).

[62] J.W. Strutt (Baron Rayleigh), *The Theory of Sound*, 1st edn (London: Macmillan & Co., 1877), p. 75 (para 78).

[63] J.C. Maxwell, 'Felici's Experiments', in *A Treatise on Electricity and Magnetism*, 3rd edn (Clarendon Press, 1891; repr. Dover Publications, 1954), pp. 182-184 (para 536), Vol 2.

[64] J.R. Carson, 'Reciprocal Theorems in Radio Communication', in *Proceedings of the Institute of Radio Engineers*, 17, no. 6 (June 1929), pp. 952-956.

[65] 'Loss Resistance' in *BR230: Admiralty Handbook of Wireless Telegraphy*, vol II, Wireless Telegraphy Theory (His Majesty's Stationery Office, 1938, repr. 1944), p. R:24.

through ionised particles in the atmosphere. For antennas, reciprocity includes the fact that the directional gain pattern is the same when transmitting or receiving.

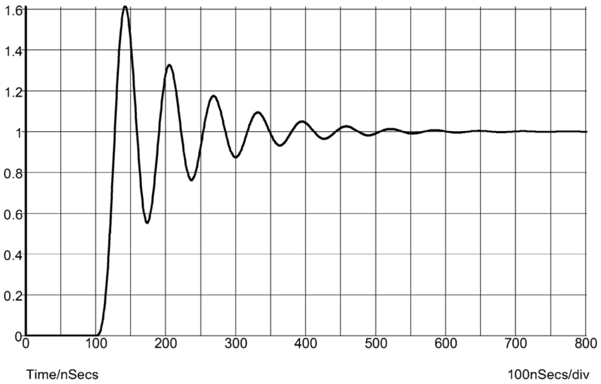
In a network represented by S-parameters, a time-invariant linear passive network will have $S_{21}= S_{12}$. This network can then be described as a *reciprocal network*.

REGULATION … has a specific legal definition: A *Statutory Regulation* is a law that must be followed. In Europe, the legislative body creates *Directives*. These directives are then made into law in each member country by *transposition* {"translation"} into a Regulation. In principle, all these Regulations are the same, having been derived from the same Directive. In practice, however, some of the fine detail is often modified in the transposition process.

The Regulations themselves can be written in loose terms. For example safety regulations tend to state that the products shall be made safe "according to good engineering practice as established in the Community". Good engineering practice is defined by what is agreed upon in recognised *standards*.

Regulation is also an electronics term for variation of output voltage with load. The transformer spec, "20% full-load regulation" means that the output voltage drops by not more than 20% from no-load to full load.

RING: If an LCR circuit, or higher order system, has insufficient damping {losses}, due to an insufficiency of series resistance, then this *under-damped* system will produce a *ring* on its pulse response.



Slightly increasing the series resistance will slow the risetime, reduce the bandwidth, reduce the overshoot, and reduce the duration of the ring. Increasing the resistance further will turn the ring into a simple single overshoot followed by a slight dip below the final settled value. Further increase can take the pulse response to *critical damping*. Still further increase makes the circuit *over-damped*, the response being ***monotonic***, but unnecessarily slow.

When tuning pulse response on a system, making the frequency of the ring faster is a step in the direction of getting a faster overall risetime in the finally tuned response.

ROGOWSKI COIL … is a form of current measuring device consisting of a toroidal coil of wire totally surrounding the conductor being probed. There is no rigid former for the toroid, so there is nothing to magnetically saturate if the current being measured has a large DC component. The Rogowski coil is also very linear as a result of not having a ferro-magnetic core.

In order to get a current reading it is necessary to amplify and integrate the output voltage, the induced voltage being the derivative of the field (by Faraday's Law of induction). In the usual form, the Rogowski coil is inside a flexible plastic tube which can be clipped together to form the toroid {ring}. It can therefore be unclipped to wrap it around a large conductor without disconnecting or disrupting the circuit being probed. This clip-together concept means that the Rogowski coil is ideal for non-intrusive in-circuit measurements. Conventional clamp-on current probes typically have jaws that are no more than a few centimetres in diameter, whereas Rogowski coils can be made much larger without incurring great additional expense.

R's and C's: Slang for resistors and capacitors. Also RC, CR, LCR networks. [L for inductance, obviously] Hyphens between the letters are optional. Usage for formal papers is not considered acceptable, but in everyday speech and writing this terminology is widely used.

RSS: Root of the **S**um of the **S**quares. Square the values, add them, then take the square root. Uncorrelated sources, such as noise sources, give a resultant power which is the sum of the individual powers of the sources. For a given load resistor, summing the squared RMS values is just like adding the individual powers.

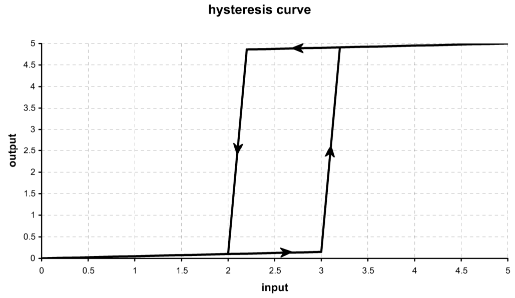
RUNT: A *runt* is a pulse of too low an amplitude, usually in a digital system. Such a pulse can be caused by incorrect terminations on a bus or a ***metastable*** state. It can also be caused by two or more outputs on a tri-state bus trying to drive the bus to complementary (opposite) logic levels at the same time. If the pulse is abnormally narrow as well, it may alternatively be called a ***glitch***. In normal usage neither a runt nor a glitch would be an acceptable signal; they would normally be indicative of a problem.

SAW: **S**urface **A**coustic **W**ave device. Ceramic resonators consist of electrodes bonded to piezoelectric ceramic material such as lead zirconium titanate (PZT). To achieve a high range of resonant frequencies, various *modes* of mechanical resonance are employed. Up to 1 MHz the modes include lengthwise vibration, area vibration and radius vibration.

The vibrational modes are related to physical dimensions within the ceramic piece, and in order to achieve the 1 MHz to 10 MHz band, the mode used is changed to thickness vibration. The 'trapped vibrations' mode can take the resonant frequency up to 100 MHz. The surface acoustic wave phenomenon is a resonance mode that is useable from around 100 MHz to just above 1 GHz.

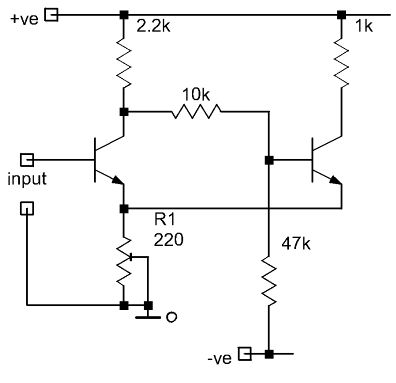
Whilst SAW devices are available up to 1 GHz and beyond, their accuracy is limited to around 0.1%. Compare this with crystal resonators: even the latest generation high frequency "mesa-" types become expensive when used directly at frequencies above 300 MHz. Crystals up to 30 MHz are the cheapest and the most accurate, achieving stabilities better than 1 ppm per year ageing and 1 ppm total shift with temperature. For higher operating frequencies, up to 40 GHz, use a *dielectric resonator oscillator*.

SCHMITT TRIGGER: A Schmitt trigger,[66] a Schmitt gate or a Schmitt input is one with internal hysteresis on the switching threshold.



This graph illustrates the hysteresis effect. The output is not solely a function of the current input; the circuit also 'remembers' a previous output condition. Hysteresis is used to eliminate problems of noise on slowly changing signals. Such noise causes multiple false transitions to be detected at logic gate inputs. If the hysteresis of the Schmitt gate is higher than the noise then once triggered, the output of the gate will stay in the "correct" logic state.

This is Schmitt's original 1938 circuit, with the exception that the original thermionic valves {tubes} have been replaced by NPN transistors; resistor values have also been added for clarification.



The circuit was devised at a time when valves were very expensive and so designers used the minimum number possible. The result is an untidy circuit in design terms. R1 is supposed to adjust the hysteresis, but also adjusts the switching thresholds and the output voltage. In fact all the component values *interact*, making a very unpleasant design to work with. The circuit also has poor immunity to temperature variation.

Circuits such as this transistorised Schmitt trigger give analog electronics a bad name, since changing one component may require you to change the values of half of the rest of the components to get the circuit to operate as intended. You should not ordinarily use this sort of 'simple' design with interactive values. A good design would have values that are *independent*. One value might set the hysteresis, another might set the switching threshold and so on. Using a circuit with a few extra

[66] O.H. Schmitt, 'A Thermionic Trigger', in *Journal of Scientific Instruments*, XV (1938), pp. 24-26.

components might actually give less design cost, shorter time to market, better maintainability, and an overall lower project cost.

SEEBECK EFFECT: Discovered by the German physicist T.Seebeck in 1821, this is the production of a voltage between a pair of series-connected junctions of dissimilar metals when one pair is at a different temperature to the other. For thermocouples, materials with a large linear *Seebeck Coefficient*, α_{AB}, are chosen so that a useful voltage is generated. $\Delta V = \alpha_{AB} \cdot \Delta T$

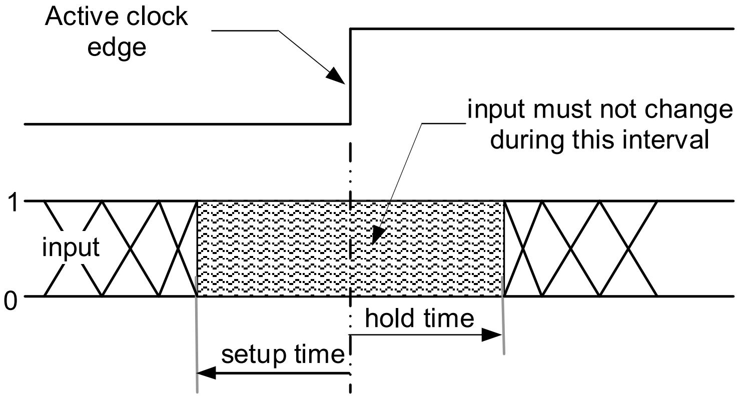
Note that the Seebeck Coefficient relates to the *pair* of metals chosen, hence the double subscript. In order to reduce unwanted thermally generated voltages for precision measurements, it is important to chose metal pairs with a low Seebeck Coefficient. See table under ***thermal EMF***.

For thermocouples, the Seebeck Coefficient is tabulated and needs to be linearised for best accuracy. When the problem is reducing the thermal EMF, the non-linearity in the Seebeck Coefficient is generally ignored.

If current is passed through these junctions then the ***Peltier Effect*** comes into play.

SETUP & HOLD TIMES : Some "designers" wire logic systems together and, since the systems "work", no further testing is done. Later, perhaps in production, they get occasional problems. Perhaps 1 in 100 times they get an inexplicable response from the system at power-up or after some other operational mode. Violation of setup and hold times is a standard reason for this type of intermittent problem.

These specs apply to clocked devices such as registers, flip-flops, latches, DACs and ADCs. There is a logic input and a clock input. For a generic clocked device there is a finite time before the active clock edge when the input data must not be changing; if the clock occurs on the rising edge, then the rising edge is the *active edge*. The time before the active clock edge when the data must be stationary is called the *setup time*. Likewise, the data must not change for some period after the active clock edge, the *hold time*. These are shown pictorially on a *timing diagram*.



Depending on the exact device used, setup and/or hold times can be zero. It is also possible for one of them to be negative. For example, a negative hold time of 1 ns would mean that the data would be allowed to change 1 ns *before* the active clock edge. Realise that setup and hold specs are worst case requirements of a logic device. If you violate the setup or hold times then any particular device may still function correctly. The problems come when the timings change as a result of temperature drifts, production spreads of skew between devices, and noise on the edge timings. Violation of setup and hold times can result in ***meta-stability*** and "random" failures in service.

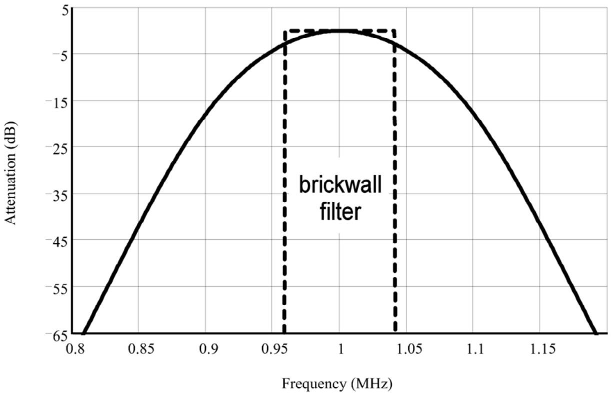
SFDR: **S**purious **F**ree **D**ynamic **R**ange … is a measure of the worst spurious signal seen on a spectrum analyser, or on an ***FFT*** of a time domain signal, when driven by a low-noise harmonically pure sinusoidal source. It is either measured in dBc, dB relative to the carrier (largest signal), or dB relative to full scale (dBFS).

$$\boxed{SFDR = 20 \times \log_{10}\left(\frac{\text{largest signal amplitude}}{\text{next largest signal amplitude}}\right) \quad \text{dBc}}$$

Both harmonic distortion and noise spikes reduce SFDR. Since spectrum analyser displays and FFTs are usually scaled in dB anyway, the SFDR can ordinarily be read directly off the display as the difference in amplitude between the largest and second largest signals on the display.

SHAPE FACTOR: For a bandpass filter, it is desirable to have steep attenuation characteristics on either side of the pass-band. This is quantified by the *shape factor*. The shape factor is the ratio of widths of pass-bands at 60 dB attenuation and 3 dB attenuation; This spec is found in spectrum analyser datasheets, but may be referred to as *selectivity*.

$$\boxed{\text{Shape Factor} = \frac{60\text{ dB bandwidth}}{3\text{ dB bandwidth}}}$$



A ***brickwall*** filter has a shape factor of 1, this being the unattainable ideal. Filters with low shape factors are needed in spectrum analysers to allow measurements close to the fundamental. If the shape factor is poor then the carrier will spread out and swamp {mask} nearby signals.

This filter curve has a shape factor of 4.4, an excellent value; 10 is very good.

It is also possible to define other shape factors for specific purposes. An example would be a 30 dB / 3 dB shape factor.

SHUNT:

1) *To shunt*, verb, is to place one component in parallel with another.
2) *Shunt*, adjective, describes a device or connection in parallel with another.
3) *A shunt*, noun, is a low value resistance placed in parallel with a component (or circuit) in order to reduce the current flow in that component. A typical application is a current shunt for a galvanometer. The shunt makes the combination into a higher current range ammeter.

SKEW … is primarily related to digital circuits, but can affect analog measurements as well. It is the time difference between two signals. For example, if you apply the same logic signal to the inputs of all six inverters in a hex-inverter IC, the outputs will not all change at the same time, even when the individual gates are loaded identically. This spec would be called “within-device skew” and is never actually mentioned for inexpensive logic gates. Only expensive clock distribution chips seem to state defined within-device skew figures.

Another form of skew is from device to device. This will be much larger than the within-device skew. This spec is obtained by looking at the max to min propagation delay variation of that particular device. You may be able to reduce the data sheet figure slightly by seeing if the limits include different capacitive loading effects; in a given application the loading should be fairly well defined.

In the analog world, you will get skew between two or more channels on a scope, even when you probe the same signal. This skew can be partly down to the scope and partly down to the probes. Some scopes have the facility to electronically de-skew the trace displayed on the screen. On other scopes, you have to probe the same point then mentally subtract the resulting time skew on all subsequent measurements.

SKIN EFFECT: The fact that RF currents flow on the outer surface of conductors has been known about for a very long time [67] and experimentally confirmed on numerous occasions. The earliest experimental evidence was necessary qualitative, but successive investigators have demonstrated the effect to greater and greater degrees of precision.

[67] J.C. Maxwell, *A Dynamical Theory of the Electromagnetic Field*, ed. by T.T. Torrance (Royal Society of London, 1864; repr., Wipf and Stock, 1982), para 115.

J. Henry used a coil of wire with a needle inside as his detector in 1877. If the needle was magnetised there had been a current through the wire. He put one coil/needle pair inside a two foot long gun barrel, and another coil/needle pair on the outside of the gun barrel. The ends of the gun barrel were sealed over with tin foil, making a good conducting seal. It was found that only the needle on the outside of the barrel was magnetised when a Leyden jar (capacitor) was discharged through the gun barrel.[68]

Hertz published his experimental results on the subject in 1889.[69] Detailed quantitative measurements to compare against the theory have also been made.[70] One set of experiments by Bjerknes (prior to 1925) measured the resistance of bare iron and copper wires. The iron wire was then plated with copper, whilst the copper wire was plated with iron. Curves of resistance against plating thickness clearly show that it is the surface region that dominates the resistance of the wire. The iron wire became progressively less resistive with increased copper plating, and the copper wire became progressively more resistive with increased iron plating.

Skin effect is often ‘explained’ in the following way: Inductance is caused by a current generating a magnetic field. A current filament in the middle of a circular conductor would be linked with more magnetic flux than a current filament in the outermost parts of this conductor. Thus the inner part of the conductor has more inductance. The increased inductance means that alternating current will find it harder to travel down the inner part of the conductor; thus most of the current will be found in the outer part of the conductor. This ‘explanation’ is **hopelessly wrong** and does not begin to explain the incredibly rapid reduction in current density that occurs in practice. To explain the effect mathematically, it is necessary to solve *Maxwell’s equations* in the metal; a task best suited to mathematicians and professors of electromagnetics.

Qualitatively, the ***Poynting vector*** approach is perhaps preferable. It is now said that the energy is in the electromagnetic field, rather than the current. The field finds it increasingly difficult to penetrate the conductor at high frequencies, the field decaying exponentially with distance from the surface. The view is taken that it is the field that creates the current, rather than the usual low frequency consideration that the current produces the field. The skin depth is therefore a ‘penetration depth’ which is also useful for evaluating the performance of shielding materials.

The current density, *J*, in a large circular conductor through which an alternating current is passing, drops off exponentially with distance, *d*, from the surface according to the relation:[71]

$$\frac{J}{J_S} = \exp\left[-(1+j)d\sqrt{\frac{\pi f \mu}{\rho}}\right],$$ where the *s* subscript is for the surface current density.

Defining the skin depth $\delta = \sqrt{\dfrac{\rho}{\pi f \mu}}$ gives $\dfrac{J}{J_S} = \exp\left[-\dfrac{(1+j)d}{\delta}\right] = \exp\left[-\dfrac{d}{\delta}\right] \times \exp\left[-j\dfrac{d}{\delta}\right]$

Only the real part of the exponential is used, giving $\boxed{\dfrac{J}{J_S} = \exp\left[-\dfrac{d}{\delta}\right] \cdot \cos\left(\dfrac{d}{\delta}\right)}$

A major reversal of current direction occurs at roughly 1.6 skin depths. At a depth of $2.36 \times \delta$, the reverse current density peaks at 6.7% of the surface current density. Note that these figures are an

[68] J.J. Fahie, 'Appendix B: Prof. Henry on High Tension Electricity', in *A History of Wireless Telegraphy*, 2nd edn (Blackwood & Sons, 1901; repr. Ayer, 2000), pp. 277-279.

[69] H. Hertz, 'On the Propagation of Electric Waves by Means of Wires.', in *Electric Waves*, trans. by D.E. Jones, originally published in Wiedemann's Annalen der Physik und Chemie 37, 1889 (Dover, 1962), pp. 160-171.

[70] A.E. Kennelly, and H.A. Affel, 'Skin-Effect Resistance Measurements of Conductors at Radio-Frequencies up to 100,000 Cycles Per Second.', in *Proceedings of the Institute of Radio Engineers*, 4 (1916), pp. 523-580.

[71] H.A Wheeler, 'Formulas for the Skin Effect', in *Proceedings of the Institute of Radio Engineers*, 30 (Sept 1942), pp. 412-424.

approximation based on the conductor radius being many times greater than the skin depth. The approximation is equivalent to neglecting the curvature of the conductor and using the plane wave solution. If the conductor radius is less than say 3× the skin depth, an accurate solution to the problem requires ***Bessel functions***.

It is convenient to consider the current density remaining at the surface density value, J_S. In this case the equivalent current would all be confined to a layer of thickness δ.

This simple concept is often used for calculating the effective resistance of a conductor at high frequencies. The cross-section of the resistive path is considered to be the perimeter of the conductor multiplied by the skin depth. Be aware, however, that this simple approximation always gives a resistance that is too low. The actual HF resistance of the conductor can easily be twice as much as this simple approximation suggests. Wide rectangular conductors, for example, are not as effective as their long cross-sectional perimeter would suggest.

For copper: $\rho = 17.2 \times 10^{-9}\ \Omega \cdot \mathrm{m}$; $\mu = 4\pi \times 10^{-7}\ \mathrm{H/m}$.

f	δ
50 Hz	9.3 mm
60 Hz	8.5 mm
10 kHz	0.66 mm
100 kHz	210 μm
1 MHz	66 μm
10 MHz	21 μm
100 MHz	6 μm
1 GHz	2 μm

This simplifies to $\delta = \dfrac{66.2}{\sqrt{f}}\ \mathrm{mm}$, for copper at room temperature.

This skin effect approximation is widely used for conductors whose radius of curvature and thickness are both several times greater than δ.

For wide rectangular strips the current is still constrained to the surface, but an additional constraint pushes the current out towards the corners of the conductor. This current redistribution is impossible to predict without detailed mathematical analysis.

Remember that the skin depth formula is all about *flux linkage*. If two or more conductors are placed right next to each other the fluxes will interact. Thus for switched-mode converter transformers, the skin effect loss is significantly increased due to the ***proximity effect***. Switched–mode transformers are therefore wound using either copper tape or with multiple paralleled wires (multi-filar winding) to minimise the copper loss.

Notice the permeability term, μ, in the skin depth formula. Now most conductors and insulators have a μ_r of around one; that is, their permeability is fairly close to that of free space. On the other hand, ferromagnetic materials containing iron, Nickel or Manganese can have μ_r values higher than 10,000. This high a μ_r will reduce the skin depth by 100×. Grain oriented 3% silicon iron (used for transformers) can have a μ_r as high as 40,000. If you make a magnetic shield out of ferromagnetic material and try to pass even low frequency AC current through it, you will find that the skin depth is very significant (200× smaller than copper and therefore ≈0.05 mm at power frequencies!).

At mm-wave frequencies (>30 GHz) the losses increase slightly faster than the skin effect formula predicts.[72] Above 50 GHz the roughness of surfaces and edges causes additional losses, amounting to tens of percent, as the surface irregularities can be as deep as the skin depth. At hundreds of gigahertz the *mean free path* of the electron approaches the skin depth and resistance increase should approach a 2/3rd power law of frequency.

When the time between electron collisions is comparable to the applied frequency, attenuation of the wave no longer occurs due to the conduction electrons. In physics books you will find the term *plasma frequency* used to describe this limiting frequency. In practice, then, it is found that the alkali metals: Cs, Rb, K, Na and Li are all transparent for far ultraviolet radiation, that is wavelengths <200 nm.

The terminology changes when shielding against X-rays and gamma rays. Instead of the skin depth, the exponential attenuation constant is called the *mean free path*. The reciprocal of the skin depth is

[72] F.J. Tischer, 'Excess Conduction Losses at Millimeter Wavelengths', in *IEEE Transactions on Microwave Theory and Techniques*, MTT-24, no. 11 (1976), pp. 853-858.

used for attenuation calculations and it is called the *macroscopic cross-section*, Σ. Shielding tables are then given in terms of the *mass attenuation coefficient*, the macroscopic cross section divided by the material density.

X-rays are specified in terms of energies in electron-volts rather than joules, because of their extremely small photon energies and extremely short wavelengths. Take visible red light, for example. At 630 nm wavelength it has a frequency of around 476,000 GHz and a photon energy of 3.15×10^{-19} joules; this energy level being more conveniently expressed as two electron-volts (2 eV).

The following table is given in terms of the "penetration depth", which is just what skin depth is. These values have been calculated from the mass attenuation coefficients given on the ***NIST*** website.[73] The density is expressed in g/cm^3, which is numerically equal to the specific gravity of the material.

Substance	Density (g/cm^3)	Penetration Depth δ vs. Photon Energy				
		1000 eV X-ray	**0.01 MeV** X-ray	**0.1 MeV** X-ray	**1 MeV** γ-ray	**10 MeV** γ-ray
Lead	11	0.17 μm	6.7 μm	0.16 mm	12 mm	18 mm
Copper	9.0	0.10 μm	5.2 μm	2.4 mm	19 mm	36 mm
Lead glass	6.2	0.33 μm	16 μm	0.38 mm	23 mm	37 mm
Iron	7.9	0.14 μm	7.4 μm	3.4 mm	21 mm	42 mm
Aluminium	2.7	3.1 μm	140 μm	22 mm	60 mm	160 mm
Concrete	2.3	1.3 μm	0.21 mm	25 mm	67 mm	191 mm
Water	1.0	2.4 μm	1.9 mm	59 mm	141 mm	451 mm
Dry air	0.0012	2.3 mm	1.6 m	54 m	131 m	406 m

Gold, platinum, iridium, osmium and uranium can all shield out gamma rays with 40% less material thickness than for lead, but the material cost increase would be prohibitive. To get a factor of ten reduction in the incident radiation, making the overly safe assumption that none is reflected, the material needs to be 2.3δ thick.

Superconductors do not follow the rules for skin depth at all. The skin depth, or penetration depth as it is called for superconductors, has been found to be constant up to around 100 GHz. Furthermore, the loss in the surface resistance increases with the square of the frequency.[74] Thus superconductors are not guaranteed to be less lossy than ordinary conductors above the gigahertz region.

SLOT ANTENNA: A slot cut in a large metal plate can act as an efficient antenna. If the slot is cut with the same shape as a known antenna the radiation resistance and field patterns can be determined from theory (Booker's extension to Babinet's principle). The complementary shape to a dipole is a simple rectangular slot. A resonant half-wave slot 0.475λ long and 0.01λ wide would have an impedance of 530 Ω if the signal were fed across the middle of the slot. To get a better match to a 50 Ω system, the feed point should be moved to 0.05λ from one end of the slot.

Note that whilst the horizontal rods in a dipole produce horizontal polarisation, a horizontal slot antenna produces *vertical* polarisation.

In general, if the terminal impedance of an antenna is Z_T, the terminal impedance of the complementary slot antenna is $Z_S = \dfrac{Z_0^2}{2\cdot Z_T}$, where $Z_0 = 377\,\Omega$ is the characteristic impedance of free space.

[73] J.H. Hubbell and S.M. Seltzer. 'Tables of X-Ray Mass Attenuation Coefficients and Mass Energy-Absorption Coefficients' (version 1.03: 1997), [Online]. Originally published as NISTIR 5632, NIST (1995).

[74] S. Ramo, J.R. Whinnery, and T.H. Duzer, 'Penetration of Electromagnetic Fields Into a Good Conductor', in *Fields and Waves in Communication Electronics*, 3rd edn (Wiley, 1994), pp. 149-155.

SLUG: Electronic slang term meaning to "kill-off" or reduce something heavily. For example: "*I slugged the bandwidth of the amplifier by putting the feedback cap up to 100 n*." It is often helpful when debugging a system to kill a possible source of a problem with a huge capacitor; just *slug it*.

Also an obsolete mass unit: 1 slug= 14.5939 kg.

SMITH CHART: The Smith Chart [75] is a graphical means of solving RF matching problems and evaluating transmission line impedances, standing wave ratios, reflection coefficients and so forth. It was developed in 1939 when computers were not available for such tasks. The mathematics behind the chart is not well explained in the original papers,[76] so you are better off seeking a specialist text book if you wish to understand and use the Smith Chart.[77] Having said that, the idea of solving design equations using charts is obsolete; such tasks should now be done directly by computer.

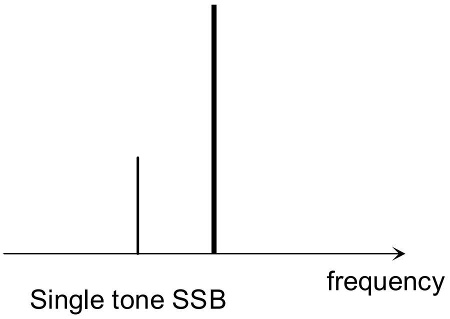
SNUBBER … is a series combination of a resistor and a capacitor used to suppress unwanted oscillations or ***rings*** on inductive circuits. This includes, but is not limited to, switch contacts, relay contacts, and inductors/transformers in switched-mode power supplies. The excess energy is dissipated in the resistor and the ring is reduced. You can regard this as lowering the Q of the parasitic inductance. Because energy is being dissipated in the snubber, the efficiency of the circuit is being reduced. This loss of efficiency can be particularly important in switched-mode power supplies. In a high efficiency circuit, it is necessary to resonate stray inductance rather than damp it, thereby maintaining the efficiency.

SPARKLE CODE … is an invalid output data word produced by an ADC which occurs relatively infrequently. The term originated in the imaging industry where such occurrences produced brief isolated spurious white dots on the screen. Reasons for these invalid codes include meta-stability in internal latches and internal propagation delay variations between comparators. This is not just a historic problem; in 2003 an experienced ADC manufacturer had to temporarily withdraw an ADC from the market due to sparkle code problems.

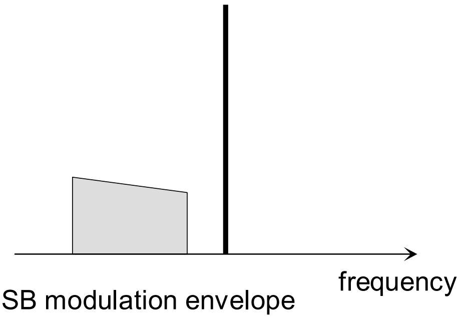
SQUEGGING … is an *unstable* oscillation. The oscillation amplitude has not been controlled properly and the amplitude builds up to the point where the amplifying device becomes unable to supply enough gain. The main oscillation amplitude therefore dies away then builds up again. The effect is to amplitude modulate the oscillation.

The term *squegging* is derived from an early form of **s**elf-**que**nching oscillator – a squegger {quench= damp, suppress}.[78]

SSB: **S**ingle **S**ide-**B**and. On a spectrum analyser display SSB would look like a large RF carrier with a single spectral line, or perhaps a band of modulation only on one side of the carrier. The horizontal axis is frequency and the vertical axis is amplitude, usually scaled in dB. Note that the shape of the modulation band, shown shaded in the diagram, is arbitrary.



Single tone SSB



SSB modulation envelope

Amplitude Modulation (AM) of an RF carrier creates a double sideband signal (see DSB). Removing

[75] P.H. Smith, 'Transmission Line Calculator', in *Electronics*, 12 (Jan 1939), pp. 29-31.

[76] P.H. Smith, 'An Improved Transmission Line Calculator', in *Electronics*, 17 (Jan 1944), pp. 133-133, +318.

[77] C. Bowick, 'The Smith Chart', in *RF Circuit Design* (Sams, 1982), pp. 75-97.

[78] *Admiralty Handbook of Wireless Telegraphy* (London: HM Stationery Office, 1938; repr., 1944), pp. F:53-F:58, Vol II, Wireless Telegraphy Theory.

one of the sidebands reduces the power required to transmit the signal, but makes the demodulation process more difficult. Removing one sideband and the carrier gives SSB-SC, single-sideband suppressed carrier. The power reduction relative to AM is greater than ×4, and the required bandwidth is halved. SSB-SC was widely used for long distance telephony up until the end of the 20th century.

STABILITY:

1) The quality of not drifting with time. This is important for components such as precision resistors.
2) The requirement for an amplifier that it should not oscillate.
3) The requirement for the gain of an amplifier to not change too rapidly with time.
4) The quality of an oscillator that means that its oscillation frequency and/or amplitude changes minimally with time. It could also mean that it has minimal ***phase noise***.
5) The quality of not shifting with temperature. This should be identified as ‘Temperature Stability’ if it is used this way. The term Temperature Coefficient - or more commonly simply TC - is preferred.

STANDARD:

1) A document produced by a government agency, an international body, or perhaps even the company you work for, that gives guidance and rules concerning some particular topic (Safety, for example).
2) A physical component, or structure of components, which is used as a local, national or international reference for measurements. This would include standard resistors, the standard metre, and standard capacitors.
3) Usually with a small “s”, a standard resistor or component can be one that is stocked by a distributor, or by your company, as a preferred part to use. Using standard “stock parts” is quicker and cheaper than using “special” (non-standard) parts.

STAR POINT: In sensitive measurement systems (say <1 mV resolution) operating at relatively low frequencies (say <100 kHz), it is usual to join all common {0 V; ground; earth} connections back to one small physical location. This is known as the *star point* and is used as the reference point for all voltages. The problem is that voltages in single-ended stages need to be referenced to something. In differential stages the reference point is the other half of the signal.

For multi-channel measurement systems it is clearly impossible to have one star point. It is therefore necessary to have individual reference points for each channel. Such a point might be the 0 V input terminal for that channel.

STATES OF MATTER: There are *six* distinct states of matter, although textbooks typically only give three.

1) Plasma
2) Gas
3) Liquid
4) Liquid crystal
5) Solid
6) Hyper-condensed Neutron star material

In this list, the molecules are progressively more bound together until. In plasma, dissociation into ions is brought about by the extreme temperatures involved, although “cold plasma” is now possible. The liquid crystal state, which has considerable importance for opto-electronic displays, is an ordered liquid and therefore is in-between the ordered array of a crystal and the complete randomness of a liquid.

Hyper-condensed matter as found in a Neutron star is thought to be >10^{13} × the density of ordinary matter and therefore necessarily has an unusual atomic structure.

STRAYS … are things like leakage resistance, coupling capacitance and mutual inductance, all of which are related to how a circuit or system is physically constructed. Guarding, shielding, or a different placement of components, can all reduce these effects. By calling an effect ‘stray’, the implication is that the effect is undesirable. Minor coupling effects which are intentional are called *loose coupling*.

The common factor in the definition of stray is that the undesirable ‘features’ are reducible by standard engineering practice. Examples of stray effects:

1) cables running close to each other, the fast edges coupling capacitively to the sensitive circuits.
2) poor choice of adjacent signals in a ribbon cable causing unexpected results.
3) cables with high currents running next to sensitive circuits and coupling magnetically.
4) high voltage tracks (or even power tracks) running near to sensitive high-impedance circuits and allowing excess leakage current to flow across unguarded PCB surfaces.

PCB effects are best considered as stray rather than parasitic; this tells the designer that they are reducible by better layout.

STUB TUNING is the action of matching an RF load to a transmission line or waveguide by using one or more short-circuited lengths of line/waveguide (stubs). This is a narrow band match and is a technique which is typically not used below 100 MHz due to the length of the lines required.

Single stub tuning consists of shunting a short-circuited length of line across the main line a specific distance from the load. The distance from the load is used to transform the load admittance into the normalised value 1+jB. The stub length is then set to give an input admittance of –jB, giving the matched impedance value of 1+j0. On a Smith chart this solution is obtained by plotting the normalised load admittance, drawing a constant VSWR circle through this point and noting where the constant VSWR circle intersects the constant conductance circle passing through the 1+j0 point. It is more convenient to use a simple computer program to solve this sort of problem. My version is *Stubby* (**www.logbook.freeserve.co.uk**)

Single stub tuning is not convenient for general purpose tuners because the variable positioning of the stub from the load is mechanically difficult. Dual or multiple stub tuners solve this problem by keeping the stubs at fixed distances from the load and adjusting the stub lengths. Each stub can only add a shunt susceptance. It is the distance between the stubs which changes the load conductance by moving around a constant VSWR circle.

For waveguides an *E-H tuner* is better as any load can be matched. This uses two sliding short-circuits, one in an E-plane arm and the other in an H-plane arm, both at the same axial position along the waveguide (see photo on the front cover).

SUB-HARMONIC: Harmonics are integer multiples of the *fundamental* frequency. For example a 1 MHz signal has a second harmonic of 2 MHz, a third harmonic of 3 MHz &c. The harmonics are also known as *overtones*: 3 MHz is the third overtone of a 1 MHz signal. The term ‘first harmonic’ would refer to the fundamental, but this expression is not used.

Sub-harmonics in complicated digital systems are often integer divisions of the fundamental. The second sub-harmonic of 1 MHz is 500 kHz. Unchanging harmonics on the sampling clock of a sample & hold, or an ADC, do not cause any sort of measurement uncertainty {error}. However, any sub-harmonics are highly undesirable because they give unwanted sampling ***jitter***.

In general, a sub-harmonic can be any less-than-unity multiple of the fundamental. In audio systems, for example, non-linearity of the loudspeaker suspension can result in weird ratios such as 5/13 of the fundamental.[79]

SUPERCONDUCTIVITY: The superconducting state is not rare. Some 26 elements become superconductors when sufficiently cooled, and a further 10 can become superconducting with the

[79] A.G.P. Peterson, 'Intermodulation Distortion', in *The General Radio Experimenter*, XXV, no. 10 (Mar 1951).

additional constraints of high pressure or thinness of sample.[80] In 1990 there were over 6000 known superconducting materials. When the temperature of a superconductor is gradually lowered, the resistance drops *abruptly* to “zero”. This “zero” has been indirectly measured as less than one part in 10^{17} of the room temperature resistance of copper. There seems to be no discernible decay of current in a superconducting loop, even over a period of more than 2 years! Thus it is thought that this may really be a true zero, despite my general protests that ‘floating point’ quantities are never zero!

Whilst the electrical resistance of Mercury drops to nothing within a few hundredths of a degree, the resistance of higher temperature superconductors can drop over one or more degrees Kelvin. Furthermore, intense magnetic fields, of say several Teslas, can spread the superconducting transition region over tens of degrees.

Of the elements, Niobium has the highest *critical temperature*, the temperature at which it becomes superconducting in the absence of a magnetic field. However, at 9.3°K this is somewhat difficult to utilise.

Too much magnetic field causes a superconductor to revert to a non-superconducting state (The *Silsbee effect*, discovered in 1916). For Niobium the critical field intensity is a maximum of 0.2 A/m and this drops to 0 as the temperature increases from absolute zero to the critical temperature. See also the ***Meissner Effect***. Niobium-Titanium alloy is a practical superconductor, allowing 10 T fields when cooled below 2°K.

Superconductors make powerful magnets possible without overheating; 25 T fields have been achieved. The current record for “high temperature” superconductors is the 165°K critical temperature of pressurised Mercury Cuprate, attained in 1994.

SUPERPOSITION THEOREM: For a linear network, the action of each individual voltage or current source is unchanged by the presence of all of the other voltage or current sources. Thus the current in any particular *branch* of a network can be evaluated by summing the currents in that branch due to each of the other sources acting independently. It is very important to note that this only applies to *linear* networks.

Superposition applies to voltage and current, but not to power, power being related to the square of voltage or current.

Superposition is one of the oldest theorems of basic physics, but its origins are unclear. It was deduced from work with waves, where it was observed that water and sound waves could pass through each other without being changed in any way.

Thévenin used the superposition theorem in 1883 to construct the ***Thévenin equivalent*** of a circuit.

TANGENTIAL NOISE … is not a type of noise, but is a measurement method described by Garuts & Samuel (1969). It should be described as ‘noise measured tangentially’ or by the ‘tangential method’. This method can be used for real-time scopes and other systems with visual display systems in order to get a semi-quantitative and (fairly) repeatable measure of the displayed noise.

On a two channel real-time scope, select *add-mode* then apply a small square wave signal to CH1, with the noise signal on CH2. Select *auto trigger*, with the trigger level well away from the signal so that the trigger is not locked on the square wave. The display should show two noisy horizontal bands on the screen, provided the square wave is bigger than the noise. Reduce the square wave amplitude so the noise bands overlap and the dark band between them has just disappeared. The amplitude of the added squarewave at this point is the measure of the noise on the waveform.

Nowadays you should just use the built in measurement functions of a digital storage oscilloscope to give ptp or RMS readings.

TC: Temperature **C**oefficient, usually expressed in ppm/°C. It is not unusual to refer to input offset TCs of opamps, the units being μV/°C. Although mathematically sloppy, such terms can be

[80] N.W. Ashcroft, and N.D. Mermin, 'Ch. 34: Superconductivity', in *Solid State Physics* (Harcourt College Publishers, 1976), pp. 726-735.

understood by their units. In scientific use there are linear temperature coefficients, quadratic temperature coefficients and so forth.

$$V(T) = V_R + \alpha \cdot T + \beta \cdot T^2 + \gamma \cdot T^3 + \ldots$$

The temperatures given are actually difference temperatures from some reference temperature such as 25°C. V_R is the value of the function at the reference temperature. To use such an equation it is therefore essential to know both the reference temperature and the temperature scale being used. Written more explicitly the equation would be:

$$V(T) = V_R + \alpha(T - T_R) + \beta(T - T_R)^2 + \gamma(T - T_R)^3 + \ldots$$

In simple engineering situations, over limited ranges of say ±15°C, it is often convenient to neglect any second or higher order terms and to just give a linear TC. The linear TC given may then be significantly different to the actual first order term, in order to allow for the missing second order and higher terms.

In reality, precision resistors have dominantly second or third order characteristics because the linear TC component has been compensated out. **T**emperature **C**oefficient of **R**esistance (TCR) is often abbreviated to TC.

THERMAL EMF … is the voltage generated when junctions between dissimilar metals are at different temperatures. It is due to the ***Seebeck Effect.*** There must always be at least two junctions.

Thermal EMFs are a definite source of errors when measuring to resolutions of 100 µV and below. They are minimised by ensuring that wire routing is done in pairs so that the junctions are kept at similar temperatures. It is also wise to use materials with low *Seebeck Coefficients* (low thermal EMFs).

Material	|Seebeck Coefficient| Relative to Copper
Copper	<0.2 µV/°C
Zinc	<0.2 µV/°C
Silver	0.3 µV/°C
Gold	0.3 µV/°C
Rhodium	0.6 µV/°C
Tin/Lead (solder)	1 µV/°C to 3 µV/°C
Aluminium	3 µV/°C
Tin	3 µV/°C
Lead	3 µV/°C
Iron	10 µV/°C
Nickel	20 µV/°C
Kovar	40 µV/°C to 75 µV/°C
Silicon	400 µV/°C
Copper Oxide	1000 µV/°C

THÉVENIN EQUIVALENT: A complicated network of components can be replaced by a simplified network consisting of a voltage source in series with a resistor.[81] The voltage source is the value of the open-circuit {unloaded} voltage of the network. The resistor value is chosen to give the same short-circuit current as the actual network. This "equivalent circuit" does not dissipate the same power as the original. It is *only* equivalent from the point of view of the load or source that it is connected to.

An LCR network cannot be replaced by an equivalent simplified LCR network for all frequencies. However, an equivalent over a limited band of frequencies is possible. If the original network consists entirely of voltage sources and capacitances, a simplified equivalent network of a single voltage source and a capacitor will work across all frequencies. Likewise an equivalent inductor and voltage generator pair can be used to replace any network of inductors and voltage sources. In general, any voltage source in series with a complex impedance { R + jX } which is representative of a more complex network would be called the Thévenin equivalent.

[81] L. Thévenin, 'Sur un Nouveau Théorème d'électricité Dynamique', *Comptes Rendus* (1883), 159-161.

Whilst the usual theoretical way of obtaining the values for the Thévenin equivalent consists of evaluating the open-circuit voltage and the short-circuit current, a real circuit will probably 'object' to being short-circuited {it will break}. In general therefore, a Thévenin equivalent will only be valid over a limited range of load or sink currents.

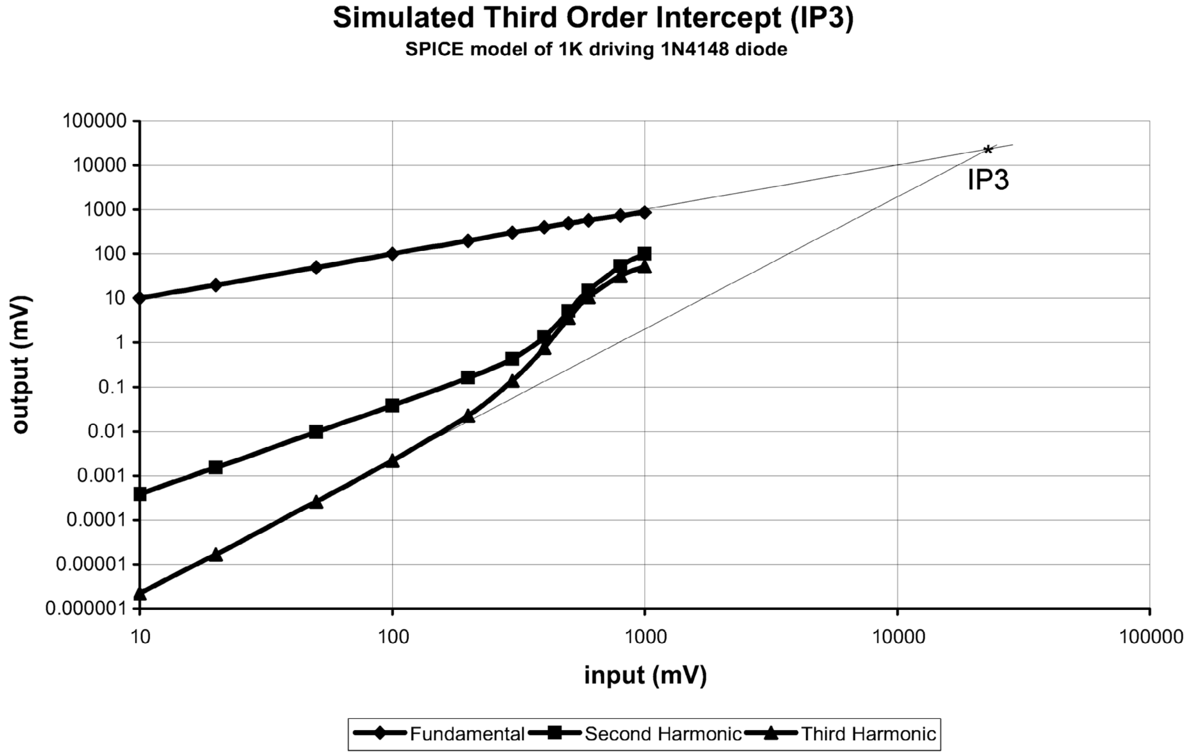
Note that the short-circuit current of a network, or its Thévenin equivalent, is not necessarily the maximum current that can be pulled from the network. Suppose that the output impedance of the network is $Z = R + jX$. The current would be a maximum when the load was an impedance of $-jX$, a resonance effect.

THIRD-ORDER INTERCEPT (IP3) … for a linear amplifier, filter, or other notionally linear device, IP3 is an extrapolated output signal level used to compare the linearity of one device to another.

In mathematics a polynomial would define a curve so that: $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots$

For a sinusoidal input: $y = a_0 + a_1 V \sin(\omega t) + a_2 V^2 \sin^2(\omega t) + a_3 V^3 \sin^3(\omega t) + \ldots$

Using trig identities, the sine cubed term can be seen to contain a triple frequency term, a *third harmonic*. This third harmonic increases as the third power of the input signal. Plotting the amplitude of the fundamental output signal against the input signal level on log-log scales ideally gives a slope of +1; doubling the input signal should double the output signal. On the same graph the third harmonic content will have a slope of +3. The third harmonic level is initially low, but as the input signal increases the third harmonic increases faster than the fundamental. The *extrapolated* output signal at which the third harmonic output level is equal to the fundamental output level is the *third order intercept*, point IP3.



In practice the device may limit or even be destroyed before this third order intercept level is reached. IP3 is given so the user can estimate how much third harmonic distortion will be present at some much lower signal level. Suppose an amplifier has a third order intercept at an output level of 15 dBm. The fundamental and third harmonics are extrapolated to be equal at this level. At an output level of 0 dBm the estimated third harmonic will be –30 dBm. The changes in amplitude, when measured in decibels, are three times larger for the third harmonic than for the fundamental.

The third order intercept is more important than the second order intercept because it is third order distortion which gives the critical third order *intermodulation products*. In fact for narrow band RF circuits, IP3 is measured by simultaneously applying two test tones which are close in both frequency and amplitude. The resulting third order intermodulation products are measured and used to calculate IP3.

THOMSON EFFECT: In 1855 the English physicist William Thomson (who later became Lord Kelvin) explained the ***Seebeck*** and ***Peltier*** effects theoretically and predicted a new one by thermodynamic principles. If a conducting rod is exposed to a temperature gradient (one end is hotter than the other) then a current flowing down the wire can cause a reversible flow of heat proportional to the temperature difference and to the current.

The close relationship between the conduction electrons in a metal and the thermal conductivity is further exemplified by the empirical Wiedemann-Franz law (1853). The ratio of thermal to electrical conductivity is directly proportional to absolute temperature for a large number of metals, the constant of proportionality also being within a few tens of percent.[82]

When routing wires into a Dewar {cryogenic cooling vessel} the massive temperature differences involved mean that significant thermoelectric voltages are generated in the wires due to the Thomson effect. For this reason it is essential to use wire from the same reel, rather than using different coloured wires. This minimises the thermoelectric mismatch.

TIN … is a metallic element which is a constituent of solder. When used as a verb, “tin a wire”, it means ‘to coat in solder’. Stranded wires can be dipped into molten solder to consolidate the strands; the wire is “tinned”. Tinning wires like this is no longer considered safe for wires inserted into screw terminal fittings however. The problem is that the solder ‘cold flows’ under the pressure of the screw and the joint becomes loose with time.

PCBs are usually *tinned* to present a good surface for soldering, although this process is also referred to as *hot air solder levelling*. Just to add to the confusion, it is possible to make electrical shielding cans from mild steel. Since it is difficult to solder to such a shield, the shield is dipped in pure tin, giving a thin tin plating. This time it really is tinned rather than being coated in solder! It should come as no surprise to learn that pure tin is an excellent coating to solder to, since solder already contains a high proportion of tin.

It is also a good idea to tin plate brass and copper parts if you wish to solder to them easily. The tin thickness only needs to be a few microns.

TRACEABLE … means you can *trace* {track; follow} back a measurement, with a defined uncertainty, through an unbroken chain of comparisons to national reference standards. This trace would be done by a ‘paper trail’. There would be a definite route that was documented for each link in the chain, each document showing the absolute uncertainty at that stage. A calibration certificate only really has traceability when issued by an externally vetted and certified calibration laboratory.

Traceable also has a meaning in non-measurement applications. In this case it means having a paper-trail for tracking a component back to its point of origin. This is essential when trying to improve reliability. This part is defective: ok, where did it come from, and are there others like it? How did it happen and how can I prevent that same error from occurring again? Without being able to track a component back to its point of origin these questions cannot be answered. It is for this reason that batch codes or date codes are often required on custom or semi-custom parts or assemblies.

If you are in a manufacturing environment for more than a few years, you will definitely encounter faulty batches of components. When this happens you will be grateful for batch codes, allowing a whole faulty batch to be removed without having to individually re-test each component.

TRACKING … is the relative matching of a particular parameter between two or more components.

[82] N.W. Ashcroft, and N.D. Mermin, 'Thermal Conductivity of a Metal', in *Solid State Physics* (Harcourt College Publishers, 1976), pp. 20-25.

Examples include:

- ☺ resistance TC matching for a pair (or set) of resistors
- ☺ matching of input offset voltage TCs between opamps in the same package
- ☺ matching between V_{BE} of two or more transistors in the same package or in the same area of a PCB.
- ☺ propagation delay matching of inverters in an IC.

The tracking value of a parameter should be less than the absolute value and there must always be at least two components involved.

Tracking is also where high-voltages arc across the surface of a PCB or other insulator. Tracking occurs at a lower voltage where there is surface contamination, such as dust or moisture. On a PCB, there is an initial voltage which causes tracking. Once an arc has crossed the gap, however, the board surface is then contaminated with carbonised deposits. This makes a second arc easier to form. The breakdown voltage can be considerably reduced from the initial value, say by a factor of 2. For this reason, spark protection gaps across the surface of PCBs are not satisfactory. The correct way to make spark protection gaps is to actually make a slot in the PCB material. There is no material to be carbonised and the arcing will occur at a more consistent voltage level.

Tracking also refers to PCB tracks {traces}. Example: “The tracking is really tight in that area.”

TRANSFER (CALIBRATION): One device is calibrated against a similar device. This could be two oscillators, two voltmeters, two sources, or two measuring devices. In order to do this calibration you need a *transfer device*. Suppose you are calibrating one DVM against another; a short-term stable voltage source is needed. Measure the voltage with the known {calibrated} DVM and then with the unknown DVM. Adjust the second DVM reading to be equal to the first, and the second is then calibrated.

Every time you make an additional transfer in a chain of transfers, you increase the uncertainty in the final measurement. The primary uncertainty in the transfer is the short term noise of each device.

TRANSFER IMPEDANCE: [Also called surface transfer impedance] … is a measure of the shielding effectiveness of a coaxial cable screen. The ideal value is 0Ω [even for cable with a characteristic impedance of 50 Ω]. An RF current is passed along the coax screen and returns through the copper tube into which the coax cable has been inserted. The voltage induced on the coaxial inner conductor is measured. Theoretically the result will be zero. In practice the construction of the screen is the limiting factor. Ordinary coaxial cable has a braided screen. This is acceptable up to perhaps a few tens of megahertz. Above this double screened cables are better. These can have a foil screen and a braid screen. At frequencies above 1 GHz a solid screen is used. These rigid or semi-rigid coax cables are essential for microwave systems.

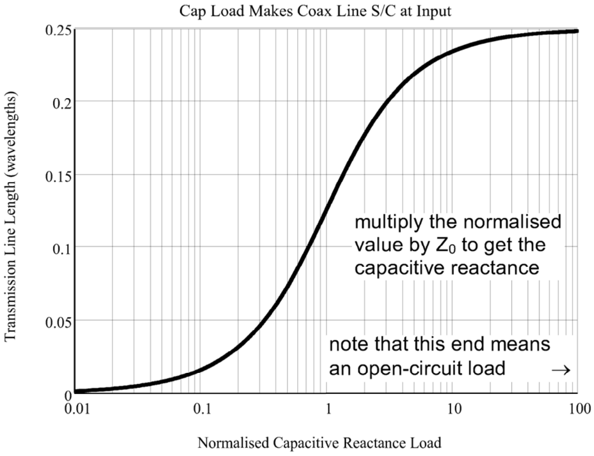
The finite value of the transfer impedance means that the cable ‘leaks’ radiation at the signal frequency. It is not unusual to get cross-talk and other interactions between coaxial cables and other circuitry at frequencies above a few hundred megahertz.

TRANSMISSION LINE: When an electrical circuit becomes longer than say 1/10th the wavelength of the signal frequency being passed through it, the phase of the signal changes significantly from one end of the connecting wire to the other. It is then necessary to consider the wire as a transmission line. When the loss in the transmission line is small enough to neglect, the input impedance of the transmission line is given by:

$$Z_{IN} = Z_0\left[\frac{Z_L\cos(\beta x) + jZ_0\sin(\beta x)}{Z_0\cos(\beta x) + jZ_L\sin(\beta x)}\right]$$

where x is the distance from the load Z_L , β is the *phase constant* for the line in rad/m, and Z_0 is the characteristic impedance of the line.

The input to the line will be effectively short-circuited when the numerator of the above expression becomes zero. This will occur for any capacitive load, but the larger the capacitive load the shorter the length of line required.



Equating the numerator to zero:

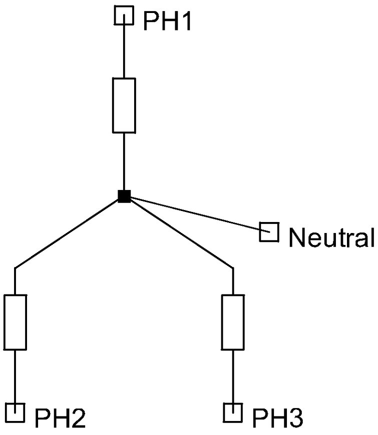
$$\frac{1}{j\omega C}\cos(\beta x) = -jZ_0 \sin(\beta x)$$

$$\therefore \beta x = \arctan\left(\frac{1}{2\pi f C Z_0}\right)$$

$$\frac{x}{\lambda} = \frac{1}{2\pi}\arctan\left(\frac{1}{2\pi f C Z_0}\right) = \frac{1}{2\pi}\arctan\left(\frac{X_C}{Z_0}\right)$$

A low-loss 50 Ω transmission line 0.125λ long with a 50 Ω capacitive load will behave like a short-circuit. Removing the capacitive load will require the line to be 0.25λ long to achieve the same effect.

TRIPLEN HARMONIC: On three-phase power systems, particular harmonics cause a disproportionate amount of trouble. These *triplen* harmonics do not ‘cancel out’ in the neutral return wire.



Consider a star-connected, balanced three-phase load. In the ideal linear case, the neutral current will be zero because the phase current returns all cancel. In the case of a non-linear load there are harmonic currents.

Since I have simplified the argument for balanced loads, the terms I_n and ϕ_n are common to all phases. The phase shift terms $\frac{2\pi n}{3}$ represent the 120° phase separation between the phases on a three phase supply.

$$I_{PH1} = \sum_n I_n \cdot \sin(n\omega t + \phi_n + 0)$$

$$I_{PH2} = \sum_n I_n \cdot \sin\left(n\omega t + \phi_n + \frac{2\pi n}{3}\right) \qquad I_{PH3} = \sum_n I_n \cdot \sin\left(n\omega t + \phi_n - \frac{2\pi n}{3}\right)$$

The sum of these three currents gives the neutral current. First expand the sine terms using a standard trig identity:

$$\sin(n\omega t + \phi_n + 0) = \sin(n\omega t + \phi_n)\cos(0) + \cos(n\omega t + \phi_n)\sin(0)$$

$$\sin\left(n\omega t + \phi_n + \frac{2\pi n}{3}\right) = \sin(n\omega t + \phi_n)\cos\left(\frac{2\pi n}{3}\right) + \cos(n\omega t + \phi_n)\sin\left(\frac{2\pi n}{3}\right)$$

$$\sin\left(n\omega t + \phi_n - \frac{2\pi n}{3}\right) = \sin(n\omega t + \phi_n)\cos\left(-\frac{2\pi n}{3}\right) + \cos(n\omega t + \phi_n)\sin\left(-\frac{2\pi n}{3}\right)$$

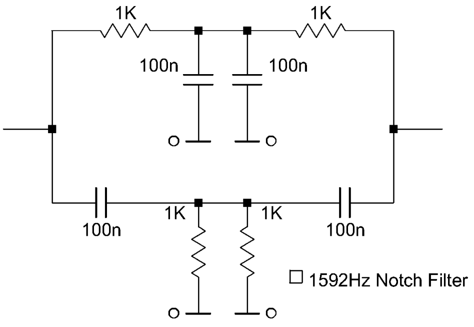
Since $\sin(0) = 0$ and $\sin(\theta) = -\sin(-\theta)$, the sum of the above terms has no $\cos(n\omega t + \phi_0)$ component. All of those terms are zero or cancel. Given that $\cos(\theta) = \cos(-\theta)$:

$$\boxed{I_{NEUTRAL} = \sum_n I_n \cdot \sin(n\omega t + \phi_n)\left[1 + 2\cdot\cos\left(\frac{2\pi n}{3}\right)\right]}$$

The square bracket evaluates to zero for all *n*, except when *n* is a multiple of three; hence the name *triple-n* or *triplen* harmonics. It is unusual to have significant quantities of even harmonics, since this means that the load is doing something different in one half-cycle compared to the next. The nuisance harmonics are therefore the odd multiples of 3, namely 3rd, 9th, 15th, 21st …

Non-linearities in transformers cause harmonic currents, but the greater problem is with rectifier circuits feeding capacitors to create low ripple DC supplies. The rectifiers conduct for such a short part of the cycle that the individual harmonic currents can be nearly as large as the fundamental. Consider the extreme example of the third harmonic being equal to the fundamental. If the fundamental current is 1 unit, the RMS phase current is $\sqrt{2}$ units, but the neutral current is 3 units; more than double the phase current!

TWIN-T FILTER …originally known as *the parallel T network*,[83] a widely published notch filter design with relatively poor performance. Its key benefit nowadays is that it can produce a passive notch with reasonable performance at audio frequencies. When dealing with harmonic distortions below −100 dBc, measurement of harmonic distortion in oscillators, amplifiers and acquisition systems becomes very difficult. Notch filtering the fundamental by >50 dB relative to the harmonics effectively gives an extra 50 dB resolution to the harmonic measuring system. In this case the Twin-T filter is used to give an 'absolute response', since its harmonic distortion cannot be measured.



The notch frequency is given by $f_N = \dfrac{1}{2\pi CR}$, where *C* is the value of each of the capacitors and *R* is the value of each of the resistors. The insertion loss at $2f_N$ is 9 dB, whilst that at $3f_N$ is 5 dB. These figures must be used to correct the second and third harmonic amplitudes when a twin-T is used to suppress the fundamental.

For this application, the resistors should ideally be 0.1% 25 ppm/°C or better. Use only NP0/C0G ceramic capacitors or film/foil capacitors. The capacitors should all be matched; get too many of the best ones you can buy and match them by measurement to better than 0.5% spread.

The resulting worst case notch depth is $\boxed{\textit{Notch Depth} \ge 64 - 20 \cdot \log_{10}(\text{sum of \% errors})\ \text{dB}}$

If one component is off by 1% the guaranteed notch depth is only 64 dB. If all 8 components are ±0.1%, and the errors combine in the worst possible way, the sum of errors would be 0.8%, giving a 66 dB notch. However a 0.1% dissipation factor in the capacitors can also reduce the notch depth to 66 dB. Increasing either of the series capacitors by 0.1% will restore the notch depth, as will decreasing the shunt resistor. A fine trim on one of the components in the twin-T will therefore ensure the maximum possible notch depth.

By unbalancing the network it is possible to get a sharper notch,[84] with up to 4 dB less insertion loss at $2f_N$. However the improvement may not justify the resulting non-integer component ratios. The impedances on the left remain unchanged. The *impedances* on the right are multiplied by a factor *y,* with *y>1.* The impedances in the centre are multiplied by $\dfrac{2y}{1+y}$. If *y* is made 5, for example, the rightmost resistor would be increased to 5K, the rightmost capacitor would be *decreased* to 20 nF, the centre resistors would be increased to 1.66667K, and the centre capacitors would be decreased to 60 nF. The frequency is calculated using the leftmost RC values.

When used for harmonic measurements by fundamental suppression, X7R and Z5U high-k ceramics

[83] H.H. Scott, 'A New Type of Selective Circuit and Some Applications', in *Proceedings of the Institute of Radio Engineers*, 26 (1938), pp. 226-235.

[84] A. Wolf, 'Note on a Parallel-T Resistance-Capacitance Network', in *Proceedings of the IRE*, 34 (1946), p. 659.

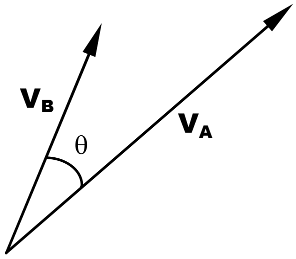
are totally useless, due to their own excessive harmonic distortion. Metallised film capacitors are also to be avoided because they are known to produce distortions up to the –90 dBc level; an amount which is very variable from manufacturer to manufacturer, and also from batch to batch.

TWO-PORT NETWORK: A network having two ports; an input port and an output port. In English, a 'port' is a gate, hole or opening through which water, air or light can pass. In the electronics context, it is an opening or connection for electricity.

The terms 'input' and 'output' on a two-port network can sometimes be misleading, as passive networks can often be operated with either end as the input. Note that there is no requirement for a DC electrical connection between the input and output ports, or indeed from any terminal to any other terminal. Such a network has also been called a *quadripole network* and a *two terminal-pair* network.

VCO: **V**oltage **C**ontrolled **O**scillator. A key element of a phase-locked loop (***PLL***), it is any oscillator whose frequency can be changed in a deterministic way by changing the voltage on a DC control line, or even changing the voltage on the power rail.

VECTOR PRODUCT (CROSS PRODUCT): The cross-product of two vectors is another vector which is perpendicular to the plane containing the original two vectors. Its magnitude is the product of the magnitudes of the two vectors and the cosine of the [smallest] angle between them. It is the product of the 'out-of–phase part' of one vector relative to the other. Note the bold type face used for vectors.



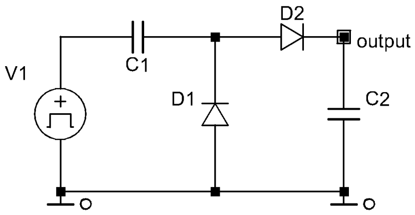
$$\mathbf{V}_C = \mathbf{V}_A \times \mathbf{V}_B = -\mathbf{V}_B \times \mathbf{V}_A$$

$$|\mathbf{V}_C| = |\mathbf{V}_A| \times |\mathbf{V}_B| \times \cos(\theta)$$

With $\mathbf{V}_A$ and $\mathbf{V}_B$ in the plane of the page, the direction of $\mathbf{V}_C$ is straight out of the page.

A related term is the *scalar product*. This is a number, a scalar quantity, whose magnitude is the product of the magnitude of the two vectors and the sine of the [smallest] angle between them. A scalar product can be thought of as the product of the 'in-phase part' of one vector relative the other. It is also known as the *dot-product* because this is the notation used when it is written.

VOLTAGE DOUBLER:



This scheme takes the ptp voltage swing on the power source (V1) and references it to 0 V. Thus a sine source of 10 V peak amplitude produces an output of 20 V, less an allowance for the forward volt-drops of the diodes.

D1 can be connected to any power rail. In that case the output voltage is the algebraic sum of the referenced power rail and the ptp input swing, less the diodes drops.

To work out approximate capacitor values, start with the output ripple voltage for the given load current. Use $I = C \cdot \frac{dV}{dt}$ rearranged to give:

$$\boxed{C_2 = I_{OUT} \cdot \frac{t_{CYCLE}}{V_{RIPPLE}} = \frac{I_{OUT}}{f \cdot V_{RIPPLE}}}$$

1 mA at 1 kHz with 100 mV ripple requires 10 μF.

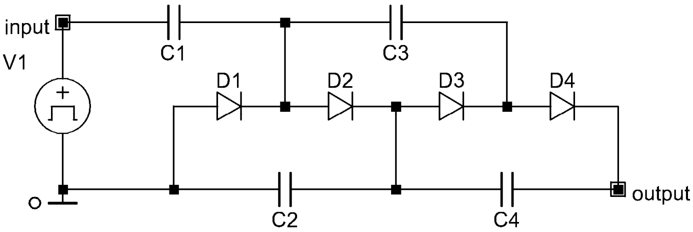
If C1 is too small then the mean value of the output voltage will be lower than expected. In fact the mean value will be related to the output frequency when C1 is small. This circuit is then a *charge pump*, but as an accurate circuit for converting frequency to voltage it is poor.

The mean output voltage will be low by the amount of discharge of C1. This is in turn governed by the amount of output current required. Every cycle the voltage change across C1 is ΔV. This gives *f* charge transfers per second. The mean current is therefore $I_{OUT} = f \cdot \Delta q = f C \cdot \Delta V$

Doubling C2 will tend to half the output ripple voltage; doubling C1 does not have such a positive effect. The droop in C1 makes the output voltage less than the peak-to-peak input voltage, but there is already a loss due to the two diode drops. Given that the diode drops will be at least 0.7 V each (for silicon diodes), there is no point in making the droop in C1 less than a few hundred millivolts.

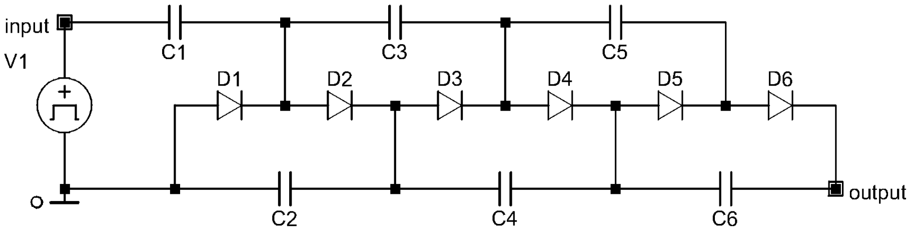
$$\boxed{C_1 = \frac{I_{OUT}}{f\cdot\left(V_{PTP}-V_{OUT}-2V_D\right)}} \qquad \text{or} \qquad C_1 \approx \frac{I_{OUT}}{f\times 0.3}$$

When using this doubler as a power supply, you should appreciate that there will be nasty voltage spikes on the output which will be made worse by poor layout. You may well need to use an LC or RC filter to minimise these switching spikes.



This is a true voltage doubler. It doubles the peak-to-peak input voltage. If the input waveform only goes from zero to positive, an improvement can be made by removing D1 and replacing C1 by a link.

Two extra capacitors and two extra diodes change the doubler to a tripler. More stages can be added using the same pattern.[85]



Out of context, the unqualified term “doubler” is ambiguous since passive diode doublers can be either “voltage doublers” or “frequency doublers”.

WAVEGUIDE: Due to the ***skin effect*** and dielectric losses, the attenuation in coaxial cables at high microwave frequencies (>20GHz) becomes unacceptably large (>1dB/m). It is found that microwave energy is more efficiently transported down hollow metal conduits with no centre conductor and no dielectric. This method has been used since 1936, such a construction being known as a *waveguide*.[86] The cross-section of the waveguide has two common shapes: circular, and rectangular with the width twice the depth. The rectangular section is the most common because it has the greatest ratio between the lowest operating *mode* and the next higher modes.

A waveguide is a low-pass filter since frequencies lower than the ***cutoff frequency*** will be severely attenuated. The idea of cutoff frequencies for waves guided through metal tubes had been predicted by Rayleigh on a purely theoretical basis some 40 years before such results were observed.[87]

Air-filled waveguides are sometimes pressurised with dry air at 4 psi to prevent the ingress of moisture. Increased moisture would cause increased attenuation and arcing at higher power levels. In fact some waveguides are run above 15 psi (≈1 atmosphere) for increased power handling capability.

[85] J.D. Cockcroft, and E.T.S. Walton, 'Experiments with High Velocity Positive Ions - (I) Further Developments in the Method of Obtaining High Velocity Positive Ions.', in Proceeding of the Royal Society of London, Series A, vol CXXXVI (1932), pp. 619-624.

[86] W.L. Barrow, 'Transmission of Electromagnetic Waves in Hollow Tubes of Metal', in *Proceedings of the Institute of Radio Engineers*, 24, no. 10 (Oct 1936), pp. 1298-1328.

[87] Lord Rayleigh, 'On the Passage of Electric Waves through Tubes, or the Vibrations of Dielectric Cylinders', in *London, Edinburgh & Dublin Philosophical Magazine and Journal of Science; Series 5*, 43, no. 261 (Feb 1897), pp. 125-132.

Rectangular waveguides can be bent without destroying their electrical characteristics provided the bend radius is greater than twice the free-space wavelength. Alternatively a right-angled bend can be used, provided the outer corner is cut off at a 45° angle and that the length of this 45° section is one quarter wavelength long; the resulting reflections then cancel, minimising the overall reflection from the sharp bend. A bend in the short side of the guide is known as an *E-plane bend*, which can be remembered by E for Easy. A bend in the long side of the guide is much harder to make. Known as an *H-plane bend*, it can be remembered as H for Hard.

The waveguide can also be twisted by 90°, provided the twist is smoothly performed over a distance of >2λ. Bends and twists are essential in microwave/mm-wave test setups to correctly orient the test pieces relative to the test bench.

WAVEGUIDE MODE $TE_{1,0}$ **:** In free space an electromagnetic wave has both the **E** field and the **H** field perpendicular to the direction of propagation {transmission; travel}, and mutually perpendicular as well; this is the **T**ransverse **E**lectro-**M**agnetic mode of propagation, TEM. TEM waves cannot travel down waveguides since the boundary conditions imposed by the conducting walls require either the electric field or the magnetic field to have a component in the direction of propagation. The direction of propagation is down the axis of the waveguide. If the electric field has no component in the direction of propagation that is a *Transverse Electric* wave, TE. In this case the magnetic field has a component in the direction of propagation.

The lowest frequency that can be sustained in a waveguide is known as the *dominant mode*. In a rectangular guide this is the $TE_{1,0}$ mode. The first suffix denotes the number of half-wavelengths across the long side of the guide, and the second gives the number of half-wavelengths across the short side. (The comma between the 1 and the 0 is optional.) Standard rectangular waveguide has the long side equal to double the short side.

Waveguide is normally operated between the $TE_{1,0}$ ***cutoff frequency*** and the $TE_{2,0}$ cutoff frequency (=$TE_{0,1}$ cutoff frequency in standard aspect ratio rectangular waveguide). Microwave components interconnected by waveguide therefore operate over a frequency range of less than one octave (2:1 frequency range). This restriction on operating frequency ensures that only one mode (the *dominant mode*) is allowed in the guide.

TM modes, **T**ransverse **M**agnetic, also exist and these use the same suffix notation.

In a resistor: $P = I^2 \cdot R$; $P = V \cdot I$; $P = \frac{V^2}{R}$

All three equations give the same power for sinusoidal excitations, but even using RMS voltages and currents for non-sinusoidal waveforms results in an error if the power is calculated from the product of current and voltage. The electric and magnetic fields are not uniform throughout the cross-section of a waveguide. Thus power flow in the guide has to be obtained from an integral. The wave impedance can then be defined in several different ways according to which pair out of voltage, current and power are used to calculate it.[88] For the $TE_{1,0}$ mode ...

$$Z_f = \frac{h}{w} \cdot \frac{377}{\sqrt{1 - \left(\frac{f_C}{f}\right)^2}}$$

$$\boxed{Z_{P,I} = \frac{\pi^2}{8} \cdot Z_f \cong 1.23 \cdot Z_f} \quad \boxed{Z_{V,I} = \frac{\pi}{2} \cdot Z_f \cong 1.57 \cdot Z_f} \quad \boxed{Z_{P,V} = 2 \cdot Z_f}$$

h is the guide height, *w* is the guide width, and f_C is the cutoff frequency in the guide.

[88] S.A. Schelkunoff, '8.21 Dominant Waves in Wave Guides of Rectangular Cross-Section' in Electromagnetic Waves (D.Van Nostrand Company, 1943), pp. 316-322.

Note that regardless of the impedance definition used, reducing the height of the guide reduces the impedance. Also note that the guide impedance changes very rapidly in the region just after cutoff. For this reason, waveguide is typically operated >20% above cutoff.

The different definitions give the standard waveguide impedance as roughly 300 Ω ± 70 Ω. If in doubt, use the $Z_{V,I}$ form. If the dielectric is not air, divide the impedance by the square root of the dielectric constant.

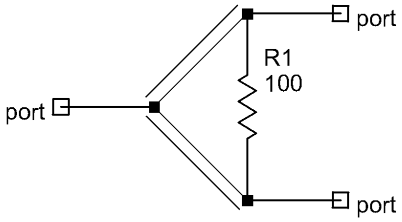
Current flows through all the walls of a waveguide with different patterns of current flow for the differing modes. In the $TE_{1,0}$ mode the current flow is only parallel to the axis of the waveguide down the middle of the two broad faces. Therefore axial cuts or splits can be made in the middle of these broad faces without creating a large insertion loss. This fact is important for making low-loss waveguide structures in two halves.

Discontinuities in a waveguide create higher order modes. If these modes are below their cutoff frequencies they cannot propagate. The result is to create localised energy storage which can be considered as inductive or capacitive elements. Such reactive elements are useful at transitions between coax and waveguide (launchers) as they can be used for tuning purposes.

WAVE IMPEDANCE: The ratio of the Electric field intensity (**E** in V/m) to the Magnetic field intensity (**H** in A/m) for an electro-magnetic wave. Far from the source, the wave impedance in free space is 377 Ω. As you get closer to a magnetic source (a changing current) the wave impedance gets progressively lower than 377 Ω. As you get closer to an electric source (a changing voltage) the wave impedance gets progressively higher than 377 Ω.

Close to the source means within 2·π wavelengths. This region is called the *Near Field*. Further away than this is called the *Far Field*. Note that in the far field, the electric and magnetic field intensities decrease *linearly* with distance (not as an inverse square law). It is the radiated power flux which decreases as an inverse square law. Since the power flux is the vector product of the **E** and **H** fields (**E**×**H** = ***Poynting vector***), both the **E** and **H** fields decrease linearly with distance.

WILKINSON COMBINER / DIVIDER … is a narrow-band RF/ microwave splitting / coupling device, based on quarter-wave transmission lines. It is difficult to get large amounts of power gain from one stage of a microwave amplifier. Therefore ingenious schemes have been devised to split the signal [losslessly], amplify it, and then recombine the paths. The Wilkinson divider is one of these.



The two sloping double-lines are quarter-wave transformers. For a 50 Ω system, the left hand port needs to see 50 Ω. Hence the two branches each need to be 100 Ω. This is achieved by making them $\lambda/4$ transmission lines having characteristic impedances of $\sqrt{50\times100} = 70.71\,\Omega$.

If power is fed in from the left then R1 does not dissipate any power because the output signals are in-phase. The right hand ports are exceptionally well matched over a wide range of frequencies. For an ideal combiner, the VSWR is better than 1.2 over a frequency span of ±52% {centre×0.48 to centre×1.52}. The left hand port has a 1.2 VSWR over the more limited frequency span of ±16%.

The exceptional matching at the right hand ports is achieved by the use of R1. A signal fed into a right hand port splits into R1 and one of the quarter wave lines. A portion of the signal sent down the quarter wave line returns to the other end of R1 through the other quarter wave line. This signal, having been through two quarter wave lines is now 180° phase shifted relative to the incoming signal and therefore subtracts from it. The value of R1 is chosen so that this cancellation is total at the centre frequency of the quarter wave transformers. A lossless reciprocal 3-port splitter can never be fully matched at all three ports (see *What are S-parameters*); the inclusion of the lossy element R1 overcomes the problem.

The original paper [89] describes an *N*-way combiner. This is more efficient than using a binary tree of divisions if a non-binary number of outputs is required. A tree division would be very inefficient, for example, if it were for 5 outputs. The other 3 outputs of the binary tree would give wasted power. For an *N*-way divider there are obviously *N* quarter wave transformers, each with an input impedance of $N \times Z_0$. Since each of these lines has an output impedance of Z_0 the impedance of the quarter wave sections is $Z_0\sqrt{N}$. The output matching resistors are all joined to a single point, with is not otherwise connected. Each of these resistors is equal to Z_0. The two way divider depicted above is therefore seen as a special case of the *N*-way divider and the two 50 Ω resistors are just joined to give a 100 Ω resistor.

WINDOW FUNCTION: An ***FFT*** is a quick method of turning a set of equally spaced time ***domain*** data points into a set of equally spaced frequency domain data points. The length of the input data set has to be an integer power of two. If it is not, then the data is either padded up to the next integer power of two by adding zeros, or truncated down to the next lower integer power of two.

The FFT 'assumes' that the data set cyclically repeats at each end of the acquired data. If the periodic waveform is not an exact sub-multiple of the measurement interval discontinuities will result. These discontinuities will appear within the frequency domain data as *spectral leakage*; single frequencies will have a finite width on the resulting spectral plot.

This undesirable leakage phenomenon is reduced by using a windowing function. This function has the effect of reducing the time domain signal towards both ends of the acquisition interval. Each acquired data point is multiplied by the window function, $w(n)$, where *n* is the position in the input array, and *N* is the length of the input array. The window function damps down the discontinuity. There is a large number of different windowing functions, with different computational requirements and different spectral responses.[90]

An important point to consider when padding with zeros and windowing is that the window must be scaled to fit the real data points and not the final padded size. The window function is designed to smoothly end at zero. If you consider padding with zeros as multiplying the window function by zero, you will see that the window function will have a sharp edge if it is applied over these padded data points. For optimal results, it is therefore essential to scale the window to fit the input data stream first, and then apply padding to get the correct length.

When the acquired data contains an exact integer number of cycles of the input waveform, there is no need to window the data. This is the 'rectangular' window function, meaning no window function at all. Some sort of windowing is essential unless the ADC clock is deliberately made an exact integer multiple of the input signal.

In general, spectral leakage decreases ***non-monotonically*** with distance from the carrier. These undesirable peaks in the spectral leakage are called *side-lobes*. One measure of the performance of a window function is therefore the first side-lobe amplitude relative to the fundamental. Although the side-lobe behaviours are shown very explicitly in textbooks and papers,[91] in practice there will be so few data points in this area that the non-monotonic behaviour will probably not be observed.

For 8-bit data the von Hann and Hamming windows are both useful, the Hamming giving a slightly better first side-lobe performance. For 16-bit data the situation is completely different. Rectangular, Gaussian and Hamming windows are useless, as they do not allow you to see the full dynamic range of the acquisition system. In this case von Hann and Blackman-Harris and are much better. The FFT plots overleaf show the effect of spectral leakage.

The flat-top window function is the most accurate for making amplitude measurements since the

[89] E.J. Wilkinson, 'An N-Way Hybrid Power Divider', in *Institute of Radio Engineers, Transactions on Microwave Theory and Techniques*, MTT-8, no. 1 (Jan 1960), pp. 116-118.

[90] F.J. Harris, 'On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform', in *Proceedings of the IEEE*, 66, no. 1 (Jan 1978), pp. 51-83.

[91] A.H. Nuttall, 'Some Windows with Very Good Sidelobe Behavior', in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29, no. 1 (Feb 1981), pp. 84-91.

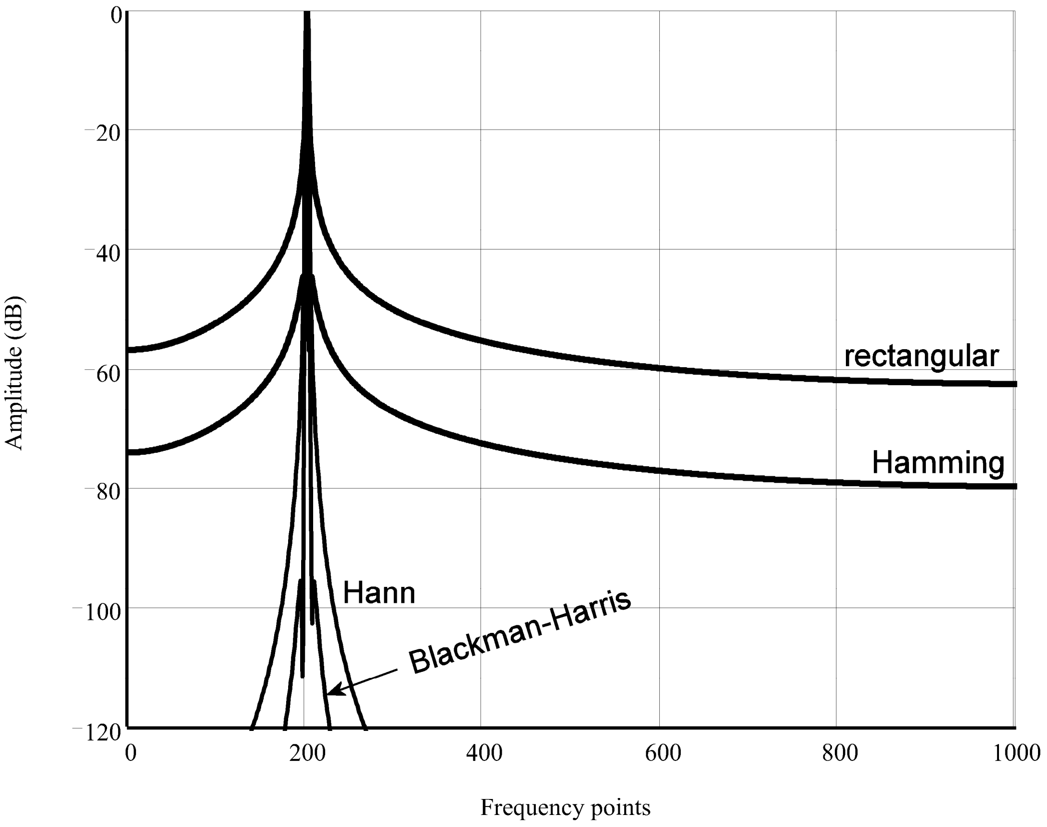
amplitude error when moving between adjacent frequency bins is minimised.

The window function called “Hanning” is an incorrect name, brought about by an early confusion between Hamming and the correct name, Hann (or von Hann).

WINDOW FUNCTION
Rectangular: $w(n) = 1; \quad 0 \le n < N$
Hann (raised cosine): $w(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right); \quad 0 \le n < N$
Hamming: $w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right); \quad 0 \le n < N$
Blackman: $w(n) = 0.42 - 0.50\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cos\left(\frac{4\pi n}{N-1}\right); \quad 0 \le \mathrm{n} < \mathrm{N}$
Flat-Top (windowed-sinc): $w(n) = \frac{1}{\left(\frac{n}{N} - \frac{1}{2}\right)}\sin\left[5\pi \cdot \left(\frac{n}{N} - \frac{1}{2}\right)\right] \times \left[0.42 - 0.50\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cos\left(\frac{4\pi n}{N-1}\right)\right]$ $0 \le \mathrm{n} < \mathrm{N}$
Blackman-Harris: $w(n) = 0.35875 - 0.48829\cos\left(\frac{2\pi n}{N-1}\right) + 0.14128\cos\left(\frac{4\pi n}{N-1}\right) \text{- } 0.01168\cos\left(\frac{6\pi \mathrm{n}}{N-1}\right)$ $0 \le \mathrm{n} < \mathrm{N}$
Force: $w(n) = \begin{cases} 0 & 0 \le n < start \\ 1 & start \le n \le stop \\ 0 & stop < n < N \end{cases}$ The start–stop region is centred on the impulse.
Exponential: $w(n) = \begin{cases} 0 & 0 \le n < start \\ \exp\left(-\tau \cdot \frac{n - start}{N - start}\right) & start \le n < N \end{cases}$ Set the start point and the damping, τ

As presented above, these formulae are not quite symmetric. In practice it is better to replace *n* by *n* – *0.5* on the right hand side in all the above formulae except the last two (force and exponential).

The force and exponential windows are primarily used for impulse testing.



Simulated FFT of a 2048 point array of equally spaced data points from a pure sinusoidal signal using various windows.

The side-lobes are just visible on the Blackman-Harris plot.

The rectangular window function is only useful when the signal is a sub-multiple of the sampling clock. A typical application is for ADC manufacturers characterising their ADCs. In this one particular case, using no window function at all gives the lowest spectral leakage.

X:

1) The unknown factor.
2) Used as a symbol for 'crystal' in abbreviations like ***OCXO*** and VCXO.
3) Used as a symbol for inductive or capacitive reactance, usually in uppercase.

$$X_C = \frac{1}{j\omega C} = \frac{-j}{\omega C}, \qquad X_L = j\omega L .$$

Note that the sign of capacitive reactance is opposite to that of inductive reactance.

4) The "don't care" state in a *truth table* input column.
5) The undefined state in a *truth table* output column.

USEFUL DATA

First Order Approximations for Small Values

It is important to quantify the error when making an approximation.

Be careful when the result is to be subtracted from unity. Consider:

$1-\sqrt{1+\delta} \approx 1-(1+0.5\delta) = -0.5\delta$, for $|\delta| \le 0.08$ the error will be up to 2%.

Approximation	Small Value	|Error|
$\sqrt{1+\delta} = 1+0.5\delta$	$\|\delta\| \le 0.085$	<0.1%
	$\|\delta\| \le 0.24$	<1.0%
$\frac{1}{\sqrt{1+\delta}} = 1-0.5\delta$	$\|\delta\| \le 0.051$	<0.1%
	$\|\delta\| \le 0.15$	<1.0%
$(1+\delta)^2 = 1+2\delta$	$\|\delta\| \le 0.03$	<0.1%
	$\|\delta\| \le 0.09$	<1.0%
$\frac{1}{(1+\delta)^2} = 1-2\delta$	$\|\delta\| \le 0.018$	<0.1%
	$\|\delta\| \le 0.056$	<1.0%
$\sin(\delta) = \delta$ $\arcsin(\delta) = \delta$	$\|\delta\| \le 0.077$	<0.1%
	$\|\delta\| \le 0.24$	<1.0%
$\cos(\delta) = 1$	$\|\delta\| \le 0.044$	<0.1%
	$\|\delta\| \le 0.14$	<1.0%
$\tan(\delta) = \delta$ $\arctan(\delta) = \delta$	$\|\delta\| \le 0.054$	<0.1%
	$\|\delta\| \le 0.17$	<1.0%
$\ln(1+\delta) = \delta$	$\|\delta\| \le 0.002$	<0.1%
	$\|\delta\| \le 0.019$	<1.0%
$\exp(\delta) = 1+\delta$	$\|\delta\| \le 0.042$	<0.1%
	$\|\delta\| \le 0.13$	<1.0%
$J_0(\delta) = 1-0.243\delta^2$ (Bessel function)	$\|\delta\| \le 0.74$	<0.1%
	$\|\delta\| \le 0.98$	<1.0%
$J_1(\delta) = \frac{\delta}{2}$ (Bessel function)	$\|\delta\| \le 0.089$	<0.1%
	$\|\delta\| \le 0.28$	<1.0%
$J_2(\delta) = \frac{\delta^2}{8}$ (Bessel function)	$\|\delta\| \le 0.10$	<0.1%
	$\|\delta\| \le 0.34$	<1.0%

The Rules of Logarithms

$\log_e(x) \equiv \ln(x)$

$\log(a \cdot b) = \log(a) + \log(b)$

$\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$

$\log(a^n) = n \cdot \log(a)$ $\qquad$ $\log_b(b) = 1$

$\log_b(1) = 0$, $\qquad$ $\log_{10}(10^x) = x$

Changing the base of logarithms:

$\ln(k) = \frac{\log_{10}(k)}{\log_{10}(e)}$, $\qquad$ $\log_{10}(k) = \frac{\ln(k)}{\ln(10)}$

$\log_2(k) = \frac{\ln(k)}{\ln(2)}$, $\qquad$ $\log_2(k) = \frac{\log_{10}(k)}{\log_{10}(2)}$

$\log_b(k) = \frac{\log_a(k)}{\log_a(b)}$, $\qquad$ $\log_b(a) = \frac{1}{\log_a(b)}$

The rule is: change the base and divide by the log (to the new base) of the old base.

The base of an exponent can be changed in a similar manner.

$$A^x = B^{x \cdot \log_B(A)}$$

… which is seen by taking logs to the base *B* of both sides. Thus an exponential decay time-constant can be expressed in terms of the *half-life*.

$$\left(\frac{1}{2}\right)^{\frac{t}{T}} = \left(\frac{1}{2}\right)^{\frac{t}{\tau \cdot \ln(2)}} = \exp\left(-\frac{t}{\tau}\right)$$

Trig Identities

$[\cos(\theta)]^2 \equiv \cos^2(\theta)$ notational style for *positive* powers of trig functions

$\arccos(x) \equiv \mathrm{acos}(\mathrm{x}) \equiv \cos^{-1}(x)$ notational styles for *inverse* trig functions.

$$\sin(-\theta) = -\sin(\theta)$$

$$\cos\,(-\theta) = \cos\,(\theta)$$

$$\sin(\theta+\phi) = \sin(\theta)\cdot\cos(\phi) + \cos(\theta)\cdot\sin(\phi)$$

$$\sin(2\theta) = 2\cdot\sin(\theta)\cdot\cos(\phi)$$

$$\sin(\theta-\phi) = \sin(\theta)\cdot\cos(\phi) - \cos(\theta)\cdot\sin(\phi)$$

$$\sin(3\theta) = 3\cdot\sin(\theta) - 4\cdot\sin^3(\theta)$$

$$\cos(\theta+\phi) = \cos(\theta)\cdot\cos(\phi) - \sin(\theta)\cdot\sin(\phi)$$

$$\cos(2\theta) = \cos^2(\theta) - \sin^2(\theta)$$

$$\cos(\theta-\phi) = \cos(\theta)\cdot\cos(\phi) + \sin(\theta)\cdot\sin(\phi)$$

$$\cos(2\theta) = 1 - 2\cdot\sin^2(\theta) = 2\cdot\cos^2(\theta) - 1$$

$$\cos^2(\theta) + \sin^2(\theta) = 1$$

$$\cos(3\theta) = 4\cdot\cos^3(\theta) - 3\cdot\cos(\theta)$$

$$\sin(\theta) + \sin(\phi) = 2\cdot\sin\left(\frac{\theta+\phi}{2}\right)\cdot\cos\left(\frac{\theta-\phi}{2}\right)$$

$$\cos(\theta) + \cos(\phi) = +2\cdot\cos\left(\frac{\theta+\phi}{2}\right)\cdot\cos\left(\frac{\theta-\phi}{2}\right)$$

$$\sin(\theta) - \sin(\phi) = 2\cdot\cos\left(\frac{\theta+\phi}{2}\right)\cdot\sin\left(\frac{\theta-\phi}{2}\right)$$

$$\cos(\theta) - \cos(\phi) = -2\cdot\sin\left(\frac{\theta+\phi}{2}\right)\cdot\sin\left(\frac{\theta-\phi}{2}\right)$$

$$\sin(\theta)\cdot\sin(\phi) = \frac{1}{2}\cdot[\cos(\theta-\phi) - \cos(\theta+\phi)]$$

$$\sin(\theta)\cdot\cos(\phi) = \frac{1}{2}\cdot[\sin(\theta+\phi) + \sin(\theta-\phi)]$$

$$\cos(\theta)\cdot\cos(\phi) = \frac{1}{2}\cdot\left[\cos(\theta+\phi) + \cos\left|\theta-\phi\right|\right]$$

Summations

$$\sum_{r=0}^{N-1} r = \frac{N}{2}(N-1) \qquad \sum_{r=0}^{N-1} r^2 = \frac{N}{6}\left(2N^2 - 3N + 1\right) \qquad \sum_{r=0}^{N-1} r^3 = \frac{N^2}{4}\left(N^2 - 2N + 1\right)$$

Power Series Expansions

$$\exp(x) = 1 + x + \frac{x^2}{2\,!} + \frac{x^3}{3\,!} + \ldots$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots \quad \text{for } |\mathrm{x}| < 1$$

$$\sin(x) = x - \frac{x^3}{3\,!} + \frac{x^5}{5\,!} - \frac{x^7}{7\,!} + \ldots \quad ;$$

$$\cos(x) = 1 - \frac{x^2}{2\,!} + \frac{x^4}{4\,!} - \frac{x^6}{6\,!} + \ldots$$

$$\sinh(x) = x + \frac{x^3}{3\,!} + \frac{x^5}{5\,!} + \frac{x^7}{7\,!} + \ldots \; ;$$

$$\cosh(x) = 1 + \frac{x^2}{2\,!} + \frac{x^4}{4\,!} + \frac{x^6}{6\,!} + \ldots$$

$$(1+x)^\alpha = 1 + \alpha\cdot x + \frac{\alpha(\alpha-1)}{2!}x^2 + \frac{\alpha(\alpha-1)(\alpha-2)}{3!}x^3 \ldots \quad \text{for arbitary } \alpha\text{, but } |\mathrm{x}| < 1$$

$$\arcsin(x) = x + \frac{1}{2}\cdot\frac{x^3}{3} + \frac{1\cdot3}{2\cdot4}\cdot\frac{x^5}{5} + \frac{1\cdot3\cdot5}{2\cdot4\cdot6}\cdot\frac{x^7}{7} + \ldots \quad \text{for } |\mathrm{x}| < 1$$

$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots \quad \text{for } |\mathrm{x}| < 1$$

Functions of a Complex Variable

$j=\sqrt{-1}$; $j^2=-1$; $\frac{1}{j}=-j$; $\sqrt{j}=\frac{1+j}{\sqrt{2}}$ imaginary number manipulations

$\exp(j\,x)=\cos(x)+j\sin(x)$ *Euler's identity*

$\sinh(j\,x)=j\sin(x)$ $\sinh(x)=-j\sin(j\,x)$

$\cosh(j\,x)=\cos(x)$ $\cosh(x)=\cos(j\,x)$

$\cos(a+jb)=\cos(a)\cdot\cosh(b)-j\cdot\sin(a)\cdot\sinh(b)$ $\sin(a+jb)=\sin(a)\cdot\cosh(b)+j\cdot\cos(a)\cdot\sinh(b)$

$\sinh(a+jb)=\sinh(a)\cdot\cos(b)-j\cdot\cosh(a)\cdot\sin(b)$ $\ln(a+jb)=\ln\left(\sqrt{a^2+b^2}\right)+j\cdot\arctan\left(\frac{\mathrm{b}}{\mathrm{a}}\right)$

$\cosh(a+jb)=\cosh(a)\cdot\cos(b)+j\cdot\sinh(a)\cdot\sin(b)$

Bessel Functions

Bessel functions are denoted by $J_n(x)$, where *n* is the *order* of the Bessel function.

$$\boxed{J_n(x)=\left(\frac{x}{2}\right)^n\cdot\left[\frac{1}{n!}-\frac{\left(\frac{x}{2}\right)^2}{1!\times(n+1)!}+\frac{\left(\frac{x}{2}\right)^4}{2!\times(n+2)!}-\frac{\left(\frac{x}{2}\right)^6}{3!\times(n+3)!}+\ldots\right]=\left(\frac{x}{2}\right)^n\cdot\sum_{k=0}^{\infty}\frac{\left(-\frac{x^2}{4}\right)^k}{k!\times(\mathrm{n}+\mathrm{k})!}}$$

Factorial notation: $n!\equiv 1\times2\times3\times4\ldots\times\mathrm{n}$; $0!=1$, $1!=1$; $2!=2$; $3!=6$; $4!=24$

Example Applications:

$$\cos(x\sin[\phi])=J_0(x)+2[J_2(x)\cdot\cos(2\phi)+J_4(x)\cdot\cos(4\phi)+\ldots]$$

$$\sin(x\sin[\phi])=2[J_1(x)\cdot\sin(\phi)+J_3(x)\cdot\sin(3\phi)+\ldots]$$

$$\cos(x\cos[\phi])=J_0(x)+2[-J_2(x)\cdot\cos(2\phi)+J_4(x)\cdot\cos(4\phi)-\ldots]$$

$$\sin(x\cos[\phi])=2[J_1(x)\cdot\cos[\phi]-J_3(x)\cdot\cos(3\phi)+\ldots]$$

Hyperbolic Functions

$\sinh(x)=\frac{\exp(x)-\exp(-x)}{2}$ $\cosh(x)=\frac{\exp(x)+\exp(-x)}{2}$

$\cosh^2(x)-\sinh^2(x)=1$ $\sinh(-x)=-\sinh(x)$ $\cosh(-x)=\cosh(x)$

$\cosh(x)+\sinh(x)=\exp(x)$ $\cosh(x)-\sinh(x)=\exp(-x)$

$\mathrm{arcsinh}(\mathrm{x})=\ln\left(x+\sqrt{x^2+1}\right)$ $\mathrm{arccosh}(\mathrm{x})=\ln\left(x+\sqrt{x^2-1}\right)$ for $x\geq1$

The Gaussian Distribution

The Gaussian distribution is tabulated with mean μ=0, and standard deviation σ=1. To use the tables, look up the value of $\frac{x-\mu}{\sigma}$.

The standard Gaussian distribution has the probability density function:

$$\boxed{p(x)=\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{x^2}{2}\right)}$$

The usual use for the Gaussian distribution requires the cumulative distribution, found by integrating the probability density function. Maths books tabulate the *single-tailed distribution*, the probability of *x* exceeding a given value. The double-tailed residual distribution, the probability of the magnitude of *x* exceeding a given value has been tabulated here.

Example: For a Gaussian distributed variable *x*, having a mean of zero and a standard deviation of 1, the chance of *x* being either greater than +5 or lesser than –5 is 0.57 ppm.

Example: For a particular large batch of resistors, having a mean value of 10.15 Ω and a standard deviation of 0.05 Ω, the probability of a chosen resistor being outside the range of 9.85 Ω < *R* < 10.45 Ω is given as 2×10^{-9} or 0.002 ppm (±6σ), assuming the resistances are normally distributed.

x	Probability of |value| > x is $1-\int_{-x}^{+x} p(z)\cdot dz$
0.500	61.71%
0.674	50.00%
1.000	31.73%
1.500	13.36%
1.645	10.00%
1.960	5.000%
2.000	4.550%
2.326	2.000%
2.500	1.242%
2.576	1.0000%
3.000	0.2700%
3.500	0.04653%
3.891	100.0ppm
4.000	63.34ppm
4.417	10.00ppm
4.500	6.795ppm
4.892	1.000ppm
5.000	5.73×10^{-7}
5.327	1.00×10^{-7}
5.500	3.80×10^{-8}
5.573	1.00×10^{-8}
6.000	1.97×10^{-9}
6.500	8.03×10^{-11}
7.000	2.56×10^{-12}
8.000	1.24×10^{-15}
10.00	1.52×10^{-23}
13.00	1.22×10^{-38}
15.00	7.34×10^{-51}

The Electromagnetic Spectrum

FREQUENCY BAND	NAME	WAVELENGTH
3-30 Hz	SAF: Sub-Audio Frequency	100,000-10,000 km
30-300 Hz	ELF: Extremely Low Frequency	10,000-1000 km
300-3000 Hz	VF: Voice Frequency	1000-100 km
3–30 kHz	VLF: Very Low Frequency	100–10 km
30–300 kHz	LF: Low Frequency, Long Wave: **LW**	10–1 km
300–3000 kHz	MF: Medium Freq, Medium Wave: **MW**	1000–100 m
3–30 MHz	**HF**: High Frequency	100–10 m
30–300 MHz	**VHF**: Very High Frequency	10–1 m
300–3000 MHz	**UHF**: Ultra High Frequency	1000–100 mm
3–30 GHz	SHF: Super High Frequency, **microwave**	100–10 mm
30–300 GHz	EHF: Extremely High Freq, **mm-wave**	10–1 mm
300 GHz–3000 GHz	sub-mm wave (**Terahertz**)	1000–100 μm
3000 GHz-20,000 GHz	**Extreme Infra Red**: (Terahertz)	100–15 μm

“Medium Wave” and “Long Wave” are designations which appear on portable radios.

Electrical Physics Notes

Two unlike charges attract each other according to Coulomb's Law $\boxed{F = \dfrac{Q_1 Q_2}{4\pi \cdot \varepsilon_0 \varepsilon_R \cdot d^2}}$

This law was experimentally determined in 1785.

A small charge placed in an electric field E experiences a force, $\mathbf{F} = q \cdot \mathbf{E}$. When the charge is moving with a velocity v and there is also a magnetic field B present, the force becomes $\mathbf{F} = q \cdot (\mathbf{E} + \mathbf{v} \times \mathbf{B})$, the *Lorentz force*.

A static electric field in space, or in matter, stores electrical energy. In a capacitor this is given by $energy = \frac{1}{2} C \cdot V^2$. Considering the energy in terms of the field $\frac{energy}{volume} = \frac{1}{2} \varepsilon_0 \varepsilon_r E^2$.

The plates on a capacitor are attracted to each other because of this electrostatic field. The electric field itself can be said to be in a state of mechanical tension. The action of one charge pulling on another is communicated by the electric field alone.

$energy = \text{force} \times \text{distance}$ and $volume = \text{area} \times \text{distance}$.

$\boxed{\dfrac{energy}{volume} = \dfrac{\text{force} \times \text{distance}}{\text{area} \times \text{distance}} = \dfrac{\text{force}}{\text{area}} = \dfrac{1}{2} \varepsilon_0 \varepsilon_r E^2}$ …the 'pulling power' of an electric field.

Two parallel wires with currents flowing in the same direction attract each other, as shown by Ampère in 1820; $\boxed{\dfrac{force}{length} = \dfrac{\mu_0 \mu_r \cdot I_1 I_2}{2\pi \cdot d^2}}$. This formula can be referred to as Ampère's Law. *Like currents attract and unlike currents repel*. This is *opposite* to the rule for charges.

Experiments show that the magnetic field around a long straight current-carrying wire is a series of concentric (coaxial) circles. These can be observed using a small magnetic compass (a plotting compass) or iron fillings spread out on a thin horizontal sheet of card through which the vertical current-carrying wire passes. *Ampère's circuital law* is then given as … $\oint_s \mathbf{H} \cdot d\mathbf{s} = I$ which says that the line integral of **H** along a closed path, *s*, is equal to the amount of current passing through the closed path. Draw any closed path in space, sum that part of the magnetic field intensity which is in the direction of the path, and do this around the whole closed path. The result is equal to the current linked to that path. The *dot product* **H**·d**s** is mathematical notation for only taking that part of **H** which is in the same direction as the small element of the path, represented by d**s**. To avoid complicated mathematics, take a simple path for *s*. In the case of a long straight current-carrying wire, for example, take a circle centred on the wire. The path length is $2\pi r$ and **H** is always in the same direction as the path. This gives $2\pi r H = I$ or $H = \frac{I}{2\pi r}$. The dimensions of **H** are *A*/m.

A static magnetic field in space, or in matter, stores energy. In an inductor: $energy = \frac{1}{2} L \cdot I^2$.

Considering the energy in terms of the field $\frac{energy}{volume} = \frac{1}{2} \mu_0 \mu_r H^2$.

The turns in an inductor are attracted to each other because of this magnetic field. The magnetic field lines themselves can therefore be said to be in a state of mechanical tension. The action of one current pulling on another is communicated by the magnetic field alone.

$\boxed{\dfrac{energy}{volume} = \dfrac{force}{area} = \dfrac{1}{2} \mu_0 \mu_r H^2}$ … the 'pulling power' of a magnetic field. Faraday first used the idea that the magnetic flux lines were under tension. He also stated that they had a mutual lateral repulsion.

When Maxwell first conceived of electromagnetic waves he thought in terms of a wave in a material substance. The electromagnetic wave was thought to propagate in a curious medium known as *the aether*. The aether wave had a kinetic (motional) aspect as well as an elastic (potential) aspect. This theory fitted in nicely with the well understood phenomena of waves in stretched strings, water &c. Thus Maxwell concluded that half the energy of an electromagnetic wave was contained in the magnetic field and half was contained in the electric field. The fields were thought to be continuously creating each other, the energy continuously shifting from magnetic to electric, and vice versa, as the wave progressed.

By around 1900 experimental searches for the aether had all proven negative and it was necessary to abandon the aether entirely. This left electromagnetic waves propagating through empty space. Furthermore the electric and magnetic fields were found to be in phase, not in quadrature, so they could not be creating each other.

In 1897 J. Larmor developed an equation for the total electromagnetic power radiated by an accelerating charge. This is a *classical theory* as the radiated power is considered to be emitted continuously; quantum theory would predict the emission of discrete photons.

In the old *electrostatic units*: for a charge *e*, accelerated by an amount *a*, with the speed of light as *c*, the radiated power *P*, was given as: $P = \frac{2}{3} \cdot \frac{e^2}{c^3} \cdot a^2$. (As given in physics books.)

In modern SI units:

$$\boxed{P = \frac{1}{6\pi\varepsilon_0} \cdot \frac{e^2}{c^3} \cdot a^2}$$

Larmor's formula was presented prior to Poincaré's presentation of "The Principle of Relativity" (1904), so 'relativistic' corrections are needed for situations where the charge is moving at a large fraction of the speed of light. These corrections are vitally important for modern particle accelerators, where the resulting radiated emissions, known as *synchrotron radiation*, can exceed megawatts!

If the acceleration is parallel to the velocity:

$$\boxed{P = \frac{1}{6\pi\varepsilon_0} \cdot \frac{e^2}{c^3} \cdot \frac{a^2}{\left[1 - \left(\frac{v}{c}\right)^2\right]^3}}$$

If the acceleration is perpendicular to the velocity:[1]

$$\boxed{P = \frac{1}{6\pi\varepsilon_0} \cdot \frac{e^2}{c^3} \cdot \frac{a^2}{\left[1 - \left(\frac{v}{c}\right)^2\right]^2}}$$

In particle accelerators, the centripetal acceleration is inversely proportional to the orbital radius. Thus the radiated power is inversely proportional to the *square* of the radius of the particle accelerator: for high energy particle accelerators, bigger is definitely better!

Deceleration is just negative acceleration, so charged particles hitting targets also produce radiation. In the case of domestic colour TV sets, the deceleration of the electron beam produces X-rays. Further information can be found by searching the WWW for *bremsstrahlung* (German for "breaking radiation" or "deceleration radiation").

When a charged particle moves through matter faster than the speed of light (in that material) *Cherenkov* radiation produced. The bluish glow around a nuclear reactor core submerged in water is an example of Cherenkov radiation.

[1] G.S. Smith, 'Ch. 6: Electromagnetic Field of a Moving Point Charge', in *An Introduction to Classical Electromagnetic Radiation* (Cambridge University Press, 1997).

Advanced Integrals and Derivatives

$$\frac{\partial}{\partial a}\int_{b}^{a} f(x)\cdot dx = f(a) \qquad \frac{\partial}{\partial a}\int_{a}^{b} f(x)\cdot dx = -f(a)$$ (where *b* is not a function of *a*)

$$J_0(x) = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \cos(x\cdot\sin[\theta])\cdot d\theta = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \cos(x\cdot\cos[\theta])\cdot d\theta$$ (Zero order Bessel function)

$$\frac{d}{dk}\cdot \boldsymbol{K}(k) = \frac{\boldsymbol{E}(k)}{k\cdot(1-k^2)} - \frac{\boldsymbol{K}(k)}{k}$$ (derivative of complete ***elliptic integral*** of the first kind)

$$\frac{d}{dk}\cdot \boldsymbol{E}(k) = \frac{1}{k}\left[\boldsymbol{E}(k) - \boldsymbol{K}(k)\right]$$ (derivative of complete elliptic integral of the second kind)

Many problems in electrostatics and electromagnetics give rise to integrals which do not have solutions in terms of elementary functions. The following set of definite integrals have solutions in terms of the standard elliptic integrals.

$$\int_0^{\pi} \frac{d\phi}{\sqrt{a\pm b\cdot\cos(\phi)}} = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{d\phi}{\sqrt{a\pm b\cdot\sin(\phi)}} = \frac{2}{\sqrt{a+b}}\boldsymbol{K}\left(\sqrt{\frac{2b}{a+b}}\right)$$

$$\int_0^{\pi} \frac{d\phi}{\sqrt{(a\pm b\cdot\cos(\phi))^3}} = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{d\phi}{\sqrt{(a\pm b\cdot\sin(\phi))^3}} = \frac{2}{(a-b)\sqrt{a+b}}\boldsymbol{E}\left(\sqrt{\frac{2b}{a+b}}\right)$$

$$\int_0^{\pi} \sqrt{a\pm b\cdot\cos(\phi)}\ d\phi = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sqrt{a\pm b\cdot\sin(\phi)}\ d\phi = 2\sqrt{a+b}\ \boldsymbol{E}\left(\sqrt{\frac{2b}{a+b}}\right) \qquad a > b > 0$$

Hypergeometric Functions

Although other hypergeometric functions do exist, unless otherwise specified assume that the Gauss Hypergeometric Function is being referred to. It is defined by:

$$F\,(a,b;c;z) \equiv {}_2F_1(a,b;c;z) \equiv 1 + \frac{a\cdot b}{c}\cdot\frac{z}{1!} + \frac{a(a+1)b\,(b+1)}{c\,(c+1)}\cdot\frac{z^2}{2\,!} + \frac{a(a+1)(a+2)b(b+1)(b+2)}{c\,(c+1)(c+2)}\cdot\frac{z^3}{3\,!} + \ldots$$

This hypergeometric function is useful for computing difficult functions such as complete ***elliptic integrals***, inverse trig functions &c. For example:

$$\arcsin(x) = x\cdot F\left(\frac{1}{2},\frac{1}{2};\frac{3}{2};x^2\right) \qquad \pi = 2\cdot F\left(1,1;\frac{3}{2};\frac{1}{2}\right)$$

$$\sin(n\cdot\arcsin(x)) = nx\cdot F\left(\frac{1+n}{2},\frac{1-n}{2};\frac{3}{2};x^2\right)$$

$$\boldsymbol{K}(k) = \frac{\pi}{2}\cdot F\left(\frac{1}{2},\frac{1}{2};1;k^2\right)$$ (complete elliptic integral of the first kind)

$$\boldsymbol{E}(k) = \frac{\pi}{2}\cdot F\left(-\frac{1}{2},\frac{1}{2};1;k^2\right)$$ (complete elliptic integral of the second kind)

APPENDIX

How much jitter is created by AM sidebands?

Create AM sidebands by multiplication of the carrier (angular) frequency ω_c, by the modulating frequency ω_m, with $\omega_c >> \omega_m$, and modulation index $m << 1$.

$$V_O = \sin(\omega_c t)\cdot[1 + m\cdot\sin(\omega_m t)] = \sin(\omega_c t) + m\cdot\sin(\omega_c t)\cdot\sin(\omega_m t)$$

The product of two sines is a standard trig identity, giving:

$$V_O = \sin(\omega_c t) + \frac{m}{2}\cdot[\cos([\omega_c - \omega_m]\cdot t) - \cos([\omega_c + \omega_m]\cdot t)]$$

This gives the large carrier with small symmetrical sidebands commonly seen on oscillators having small amounts of systematic noise. Because ω_m is so much lower than ω_c you only have to wait for one cycle of ω_m to see the maximum peak-to-peak jitter on the zero crossings of the output voltage.

With $V_O = \sin(\omega_c t)\cdot[1 + m\cdot\sin(\omega_m t)]$, it is clear that V_0 is always zero when the sine term outside of the square brackets is zero. The modulating voltage can have *no effect* on the zero crossing points and hence no jitter! However, introducing a small voltage offset on the sampling point gives a more realistic answer. $V_{OFFSET} \equiv k = \sin(\omega_c t)\cdot[1 + m\cdot\sin(\omega_m t)]$

Since ω_m is changing slowly relative to ω_c, the sine term cannot change much between one cycle of the carrier and the next. The biggest change over this period being from $-\sin\left(\frac{\omega_m}{2f_c}\right)$ to $+\sin\left(\frac{\omega_m}{2f_c}\right)$.

Since $f_c >> f_m$, $\sin\left(\frac{\omega_m}{2f_c}\right) = \sin\left(\pi\cdot\frac{f_m}{f_c}\right) \approx \pi\cdot\frac{f_m}{f_c}$

For the first crossing of the peak jitter point, $k = \sin(\omega_c t_1)\cdot\left[1 - m\pi\cdot\frac{f_m}{f_c}\right]$

$t_1 = \frac{1}{\omega_c}\cdot\arcsin\left(\frac{k}{1 - m\pi\frac{f_m}{f_c}}\right) \approx \frac{k}{\omega_c}\cdot\left(1 + m\pi\frac{f_m}{f_c}\right)$, then nearly one cycle later $t_2 \approx \frac{k}{\omega_c}\cdot\left(1 - m\pi\frac{f_m}{f_c}\right)$,

giving a peak jitter of $\delta t_{pk} = t_1 - t_2 = \frac{2k}{\omega_c}\cdot m\pi\frac{f_m}{f_c} = \frac{2\pi\cdot km}{2\pi f_c}\cdot\frac{f_m}{f_c} = km\frac{f_m}{f_c^2}$

Since the jitter is sinusoidal, the peak to RMS ratio is known. $\boxed{\delta t_{RMS} = \frac{k\cdot m}{\sqrt{2}}\cdot\frac{f_m}{f_c^2}}$

The jitter is directly proportional to the sideband amplitude. The jitter is also directly proportional to the sampling offset *k*. Remember that the sideband amplitudes are each *m*/2.

To get the best performance from any sampling system it is clear that every effort should be made to make *k* as close as possible to 0. In other words the sampling point should be in the middle of the sinusoidal clock signal. Furthermore, sidebands close to the carrier are not nearly as bad as sidebands far from the carrier.

How much jitter is created by FM sidebands?

Consider the phase modulated signal … $V_O = \sin(\omega_c t + \theta \cdot \sin[\omega_m t])$, where θ is the peak phase deviation, also known as the *modulation index β*. Using a trig expansion, $V_O = \sin(\omega_c t)\cdot\cos(\theta \sin[\omega_m t]) + \cos(\omega_c t)\cdot \sin(\theta \sin[\omega_m t])$. When θ is sufficiently small, an approximation brings the modulation terms outside of the enclosing trig functions $V_O \approx \sin(\omega_c t) + \theta\cdot\cos(\omega_c t)\cdot\sin(\omega_m t)$.

For larger θ use Bessel functions to expand the trig functions; higher order sidebands then appear.

When θ remains small … $\boxed{V_O \approx \sin(\omega_c t) + \frac{\theta}{2}[\sin([\omega_c + \omega_m]\cdot t) + \sin([\omega_c - \omega_m]\cdot t)]}$

The sideband amplitude is therefore approximately $20\cdot\log_{10}\left(\frac{\theta}{2}\right)$ dBc .

Looking at the original function, $V_O = \sin(\omega_c t + \theta \cdot \sin[\omega_m t])$, the zero crossing points occur when the argument of the outer sine function is an integer multiple of 2π radians. However, since $\omega_m \ll \omega_c$ the modulated phase has very little opportunity to modify the phase between successive zero crossing points of the carrier. Evidently the largest phase change that can be produced by the modulating signal will occur at points of the greatest slope of the inner sine function, namely around its zero crossing points. Thus the largest phase change over one cycle of the carrier will be from $\theta\cdot\sin\left(-\frac{\omega_m}{2f_c}\right)$ to $\theta\cdot\sin\left(\frac{\omega_m}{2f_c}\right)$, a change of $2\cdot\theta\cdot\sin\left(\pi\cdot\frac{f_m}{f_c}\right) \approx 2\pi\cdot\theta\cdot\frac{f_m}{f_c}$.

Equating the outer sine function to $2\pi n$ radians gives the zero crossing points.

$$2\pi f_c(T_c + \delta t) + 2\pi\cdot\theta\cdot\frac{f_m}{f_c} = n \times 2\pi \quad\rightarrow\quad \delta t_{pk} = -\theta\cdot\frac{f_m}{f_c^2} \rightarrow \quad \boxed{\delta t_{RMS} = \frac{\theta}{\sqrt{2}}\cdot\frac{f_m}{f_c^2}}$$

A pair of sidebands at a large offset frequency produces more jitter than an equal amplitude pair of sidebands closer in. Suppose you only see two sidebands on a spectrum analyser. Large sidebands, > –25 dBc, must be due to amplitude modulation (AM). Small sidebands could be AM, but could instead be phase or frequency modulated.

Angle modulated sidebands at $\omega_c \pm 2\omega_m$ have an amplitude of $2\cdot J_2(\theta^2) \approx \frac{\theta^2}{4}$, using the second order Bessel function approximation. If the spectrum analyser noise floor is –80 dBc, and if the second order sidebands are just hiding in the noise floor at –80 dBc then,

$$20\times\log_{10}\left(\frac{\theta^2}{4}\right) = -80 \quad\rightarrow\quad 40\times\log_{10}\left(\frac{\theta}{\sqrt{4}}\right) = -80 \quad\rightarrow\quad 20\times\log_{10}\left(\frac{\theta}{2}\right) = -40\,\text{dBc}$$

If the first order sidebands are below –40 dBc, any second order sidebands may not be visible. Therefore it may not be obvious if a signal being viewed on a spectrum analyser is angle modulated or amplitude modulated.

For $V_O = \sin(\omega_c t + \theta\cdot\sin[\omega_m t])$, the argument of the sine function is an angle, $\phi_i = \omega_c t + \theta\cdot\sin(\omega_m t)$, the time derivative of which is the instantaneous angular frequency $\omega_i = \frac{d\phi}{dt} = \omega_c + \theta\cdot\omega_m\cdot\cos(\omega_m t)$. $\therefore \omega_i = 2\pi[f_c + \theta\cdot f_m\cdot\cos(\omega_m t)]$. The peak frequency deviation from the carrier is seen to be the peak phase deviation times the modulation frequency, $\Delta f = \theta\cdot f_m$. In communication text books the peak phase deviation is referred to as the *modulation index*, $\beta = \frac{\Delta f}{f_m}$. When β is small, the system is referred to as narrow band FM (NBFM).

What is the basis of a phasor diagram?

Adding one sinusoidal current to another, each having different amplitudes and some arbitrary phase difference, results in yet another sinusoidal current having a new amplitude and phase. This is a key fact used throughout AC circuit theory. Expressing this idea mathematically:

$$\boxed{\sin(\omega\cdot t)+A\cdot\sin(\omega\cdot t+\phi)=B\cdot\sin(\omega\cdot t+\theta)}$$

Expand the left side using a standard trig identity:

$$\begin{aligned}\sin(\omega\cdot t)+A\cdot\sin(\omega\cdot t+\phi)&=\sin(\omega\cdot t)+A\cdot[\sin(\omega\cdot t)\cdot\cos(\phi)+\cos(\omega\cdot t)\cdot\sin(\phi)]\\&=\sin(\omega\cdot t)\cdot[1+A\cdot\cos(\phi)]+\cos(\omega\cdot t)\cdot A\cdot\sin(\phi)\end{aligned}$$

Expand the right hand side of the original equation using the same trig identity:

$$B\cdot\sin(\omega\cdot t+\theta)=B\cdot[\sin(\omega\cdot t)\cdot\cos(\theta)+\cos(\omega\cdot t)\cdot\sin(\theta)]$$

Comparing the expansions gives a pair of simultaneous equations:

$\cos(\omega\cdot t)$ terms → $B\cdot\sin(\theta)=A\cdot\sin(\phi)$

$\sin(\omega\cdot t)$ terms → $B\cdot\cos(\theta)=1+A\cdot\cos(\phi)$

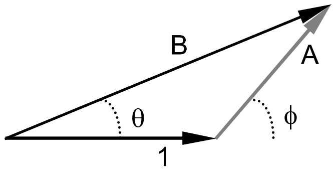
Square and add these simultaneous equations to get:

$$B^2\left(\cos^2(\theta)+\sin^2(\theta)\right)=[1+A\cdot\cos(\phi)]^2+A^2\cdot\sin^2(\phi)$$

$$\therefore B^2=1+2\cdot A\cdot\cos(\phi)+A^2\cdot\cos^2(\phi)+A^2\cdot\sin^2(\phi)$$

$$\therefore B=\sqrt{1+2A\cdot\cos(\phi)+A^2}$$

The ratio of the simultaneous equations gives $\tan(\theta)=\dfrac{A\cdot\sin(\phi)}{1+A\cdot\cos(\phi)}$



The trig equations are seen to agree with a vector diagram.

The original idea of adding two sinusoids to give a third has been proved, but the mathematical manipulations are seen to be rather unpleasant.

Rather than using sine and cosine functions, it is more convenient to use the complex form, $\exp(j\omega t)=\cos(\omega\cdot t)+j\cdot\sin(\omega\cdot t)$, Euler's Identity. This function gives a line which is rotating in the complex plane. 'Looking' at this rotating line once per cycle makes the line appear stationary and it is then called a *phasor* (historically it has also been called a complexor). The phasor can be combined with other phasors using the rules of vectors. The result is a *phasor diagram*; do not call it a 'vector diagram' because although phasors obey the rules of vectors, they have to be understood to be a snap-shot of a rotating system.

Phasors rotate anticlockwise in the complex plane. Use the mnemonic CIVIL to remember that in a capacitor (C) the current (I) leads the voltage (V), whereas the voltage (V) leads the current (I) in an inductor (L).

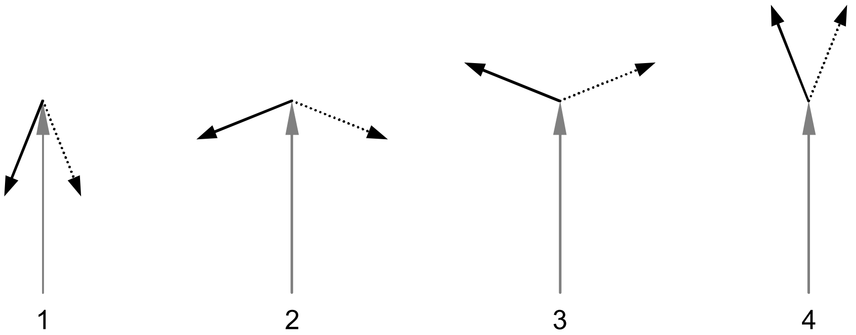
How do you draw a phasor diagram showing different frequencies?

A phasor diagram ordinarily shows phase differences and amplitudes of sinusoids of the same frequency. The phasors can then be added graphically, using the rules of vectors. The phasor diagram is a snapshot in time of a system where the phasors are considered to be rotating anti-clockwise as time advances.

Consider a phasor diagram for a signal with amplitude modulation (AM):

$$V_O = \sin(\omega_c t) + \frac{m}{2}\cdot\left[\cos\left(\left[\omega_c - \omega_m\right]\cdot t\right) - \cos\left(\left[\omega_c + \omega_m\right]\cdot t\right)\right]$$

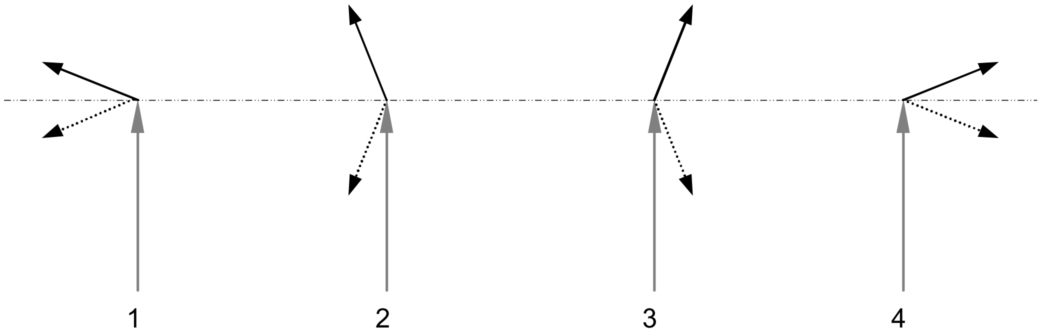
If the *modulation index*, *m*, is small and the modulation frequency $\omega_m << \omega_c$, the signal looks quite sinusoidal and the diagram is a reasonable representation. As with all phasor diagrams, the phasors are rotating but appear stationary, as if viewed by a stroboscope. Hence on the phasor diagram of the AM situation, the carrier is taken as the reference frequency and is therefore stationary. The two sidebands are now seen to be rotating on the diagram at a speed of ω_m rad/s ($\omega_m/2\pi$ rev/s) in opposite directions. The dotted phasor is rotating anti-clockwise and it is therefore the upper sideband (higher frequency).



The sidebands have been exaggerated in size relative to the carrier for the purpose of illustration. Notice that the resultant of the three phasors is in the same direction as the carrier, but cyclically changing in amplitude (hence amplitude modulation).

For phase modulation the phasor diagram is quite different:

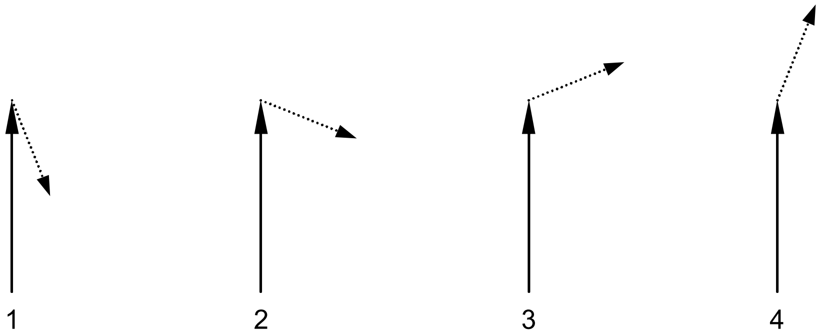
$$V_O = \sin(\omega_c t) + \frac{\theta}{2}\left[\sin\left(\left[\omega_c + \omega_m\right]\cdot t\right) + \sin\left(\left[\omega_c - \omega_m\right]\cdot t\right)\right]$$



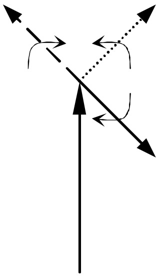
The head of the resultant phasor always lies on the horizontal line. It therefore changes its phase relative to the carrier (hence phase modulation). The resultant amplitude is the projection onto the vertical [carrier] axis and is therefore constant.

Why is single-sideband noise half AM and half PM?

First draw a phasor diagram of a single-sideband (***SSB***), single-tone signal.



This sequence of diagrams represents an upper sideband tone because the phasor is rotating anti-clockwise. Comparing this diagram with the diagrams for AM and PM it seems as if there are both AM and FM components to the modulation. To prove this, 'invent' two lower sidebands; these two invented 'virtual' sidebands are equal in magnitude, but opposite in phase to each other, therefore cancelling with each other at all times.



The dotted and dashed pair together make a pure AM signal. The dotted and plain pair together make a pure PM signal. Thus if the virtual lower sidebands are each half the amplitude of the original SSB signal, the resultant phasor is the same and the case is proved that an SSB tone can be represented as half AM and half PM.

The situation is no different for an SSB noise waveform. In reality, noise occurs on both sides of the carrier. The point is that the noise is not necessarily symmetric about the carrier frequency.

Why can you convert roughly from dB to % by multiplying by 10?

This rule is only useful for small dB amounts, not more than 3 dB, and is *only suitable for mental arithmetic.*

$$\text{voltage ratio in decibels} \equiv 20 \cdot \log_{10}(V_1/V_2)\,\text{dB}$$

Write $\frac{V_1}{V_2} = 1 + \delta$; the voltage ratio has been expressed as a per-unit difference, δ.

$$\log_{10}\left(\frac{V_1}{V_2}\right) = \log_{10}(1+\delta) = \frac{\ln(1+\delta)}{\ln(10)} = \frac{\delta - \frac{\delta^2}{2} + \frac{\delta^3}{3} - \dots}{\ln(10)} \approx \frac{\delta}{\ln(10)} \quad \text{... for small } \delta \text{ .}$$

dB value	Approx.	Accurate value
+3 dB	+30%	+41%
+2 dB	+20%	+26%
+1 dB	+10%	+12%
+0.1 dB	+1%	+1.2%
+0.01 dB	+0.1%	+0.12%
–0.01 dB	–0.1%	–0.12%
–0.1 dB	–1%	–1.2%
–1 dB	–10%	–11%
–2 dB	–20%	–21%
–3 dB	–30%	–29%

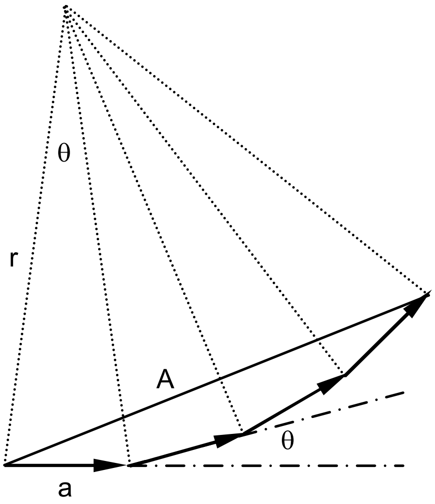
Ten times the dB value is

$$10 \cdot 20 \cdot \log_{10}\left(\frac{V_1}{V_2}\right) \approx \frac{200}{\ln(10)} \cdot \delta = 87\delta \ .$$

Multiplying δ by 100 makes it a percentage difference. This gives a relatively poor approximation. Nevertheless people do still use this result for mental arithmetic.

How is the bandwidth of a "boxcar" average digital filter calculated?

One of the simplest digital filters is achieved by taking the mean of *N* successive data points to produce one filtered data point. This type of filter is known as a *boxcar average*. The simplicity of the filter means that it can easily be implemented in hardware as an accumulation of acquired data points. Although digital filter textbooks solve the frequency response problem using complex exponentials and impulse response transformations, that method gives incorrect results due to the *Gibbs phenomenon*. The technique presented here is adapted from a physics text book solution to a diffraction problem.[1]



In this diagram, the little phasors show the phase difference between the successive points in the block to be averaged. In this example 4 points of length *a* are summed to produce a resultant of length *A*. In order to do this summation an isosceles triangle is constructed on each phasor such that the angle of phase shift, θ, between successive points is duplicated by the small angle at the top of the triangle. The angles are made equal by a suitable choice of the sides, *r*, of the triangles.

Choose $r\cdot\sin(\theta/2)=a/2$, giving the overall result implicitly as: $r\cdot\sin(N\theta/2)=A/2$.

Dividing the first equation by the second to eliminate *r* gives: $\frac{a}{A}=\frac{\sin(\theta/2)}{\sin(N\theta/2)}$.

In practice *A* is divided by the number of samples to get the mean, giving a transfer function of $\frac{A/N}{a}=\frac{\sin(N\theta/2)}{N\cdot\sin(\theta/2)}$, which equals $\frac{1}{\sqrt{2}}$ at the 3 dB bandwidth point (by definition). Solving this equation for θ is not possible analytically. For large *N*, however, it is clear that θ becomes small enough for the small-angle approximation to the sine to be used, resulting in $\frac{\sin(N\theta/2)}{N\theta/2}=\frac{1}{\sqrt{2}}$. Theta is then defined in terms of an inverse sinc function. The 3 dB bandwidth *B* is then given by: $B=\frac{F_S}{2\pi}\cdot\theta$, where F_S is the sampling rate of the original data points.

An excellent approximation (<1% error) for $N\geq 5$ is:

$$\boxed{B\approx\frac{F_S}{N\pi}\cdot\mathrm{arcsinc}\left(\frac{1}{\sqrt{2}}\right)=\frac{F_S}{N}\times 0.44295\approx 0.443\times F_{OUT}}$$

(The approximate and exact solutions are graphically compared in Chapter 11.)

arcsinc is not a standard function, but can be evaluated numerically. F_{OUT} is the output data rate of the averaged blocks of *N* input data points. The formula assumes that the input data points are not aliased and that the analog bandwidth is considerably higher than the sample rate.

[1] R.P. Feynman, R.B. Leighton, and M.L. Sands, '30 Diffraction', in *The Feynman Lectures on Physics* (Addison-Wesley, 1963; repr., 1989), Vol 1.

Why are tolerances added?

If two cascaded gain sections have gains of G_1 and G_2, and the per-unit tolerances on these gains are ±δ and ±ε respectively, then the overall gain, G, is given exactly by:

$$G = G_1 \times (1 \pm \delta) \times G_2 \times (1 \pm \varepsilon) = G_1 \times G_2 \times (1 \pm \delta \pm \varepsilon \pm \delta \cdot \varepsilon)$$

The per-unit tolerance of G_1 is also known as its *relative error*; $\delta = \frac{\Delta G_1}{G_1}$. When tolerances are added to find the 'worst case' result, the product of the two per-unit tolerances is being neglected as it is relatively less important. A little table shows how good this approximation is. Historically the tolerance was referred to as the 'error', although it was intended to be an error *band* within which the actual value would fall. More modern usage is to call this error band the *uncertainty* of the value. Percentages have been used rather than per-unit values to avoid too many zeroes.

δ	ε	Actual Uncertainty	Approximation
±1%	±5%	+6.05%, –5.95%	±6%
±2%	±2%	+4.04%, –3.96%	±4%
±5%	±5%	+10.25%, –9.75%	±10%
±5%	±10%	+15.5%, –14.5%	±15%
±10%	±10%	+21.0%, –19.0%	±20%
±1%	±20%	+21.2%, –20.8%	±21%
±2%	±20%	+22.4%, –21.6%	±22%
±5%	±20%	+26%, –24%	±25%
±10%	±20%	+32%, –28%	±30%
±20%	±20%	+44%, –36%	±40%
±30%	±30%	+69%, –51%	±60%
±40%	±40%	+96%, –64%	±80%

Tolerances below 10% work well for general calculations. Above this they get a bit awkward. For example a +10% tolerance 'underneath' converts to −10% tolerance 'on top' because:

$$\frac{1}{1.10} = 0.909 \approx 0.9$$

This rule gets progressively less useful as tolerances greater than 10% are used.

$$\frac{1}{1.50} = 0.667 >> 0.5$$

You should learn to think in ***per-unit*** form and be able to immediately convert 5% to either 1.05 or 0.95 per-unit multipliers. You can then evaluate 20% and 30% cascaded gains to:

$$1.20 \times 1.30 = 1.56 \rightarrow +56\%$$

$$0.80 \times 0.70 = 0.56 \rightarrow -44\%$$

This rule for gains is actually a small part of a much larger rule. Suppose an output voltage V_O is dependant on a network of resistors R_1, R_2 &c. The per-unit error in each resistor affects the per-unit error in the output voltage by a scaling factor, the scaling factor being determined either experimentally or mathematically. This scaling factor is actually the partial derivative of the output function with respect to the resistor being considered. So, for example, if a 5% change in a resistor causes a 2% change in the output voltage, all other variables being left constant, the partial derivative (scaling factor) is 0.4.

$$\frac{\Delta V_O}{V_O} \leq \frac{\Delta R_1}{R_1} \cdot \frac{\partial V_O}{\partial R_1} + \frac{\Delta R_2}{R_2} \cdot \frac{\partial V_O}{\partial R_2} + \ldots$$

This uncertainty relation has the same limit of applicability as the earlier example, namely that the individual uncertainties should be less than 10%. Of course the output variable could be a current, gain, or any other quantity. Also, the relative error terms on the right hand side could be due to error sources other than just resistors.

These partial derivative factors are also known as *sensitivity factors* or *weighting factors*. It is possible, but less usual, for these factors to be greater than unity.

Why are risetimes added as Root Sum of Squares?

Consider the normalised single-pole frequency domain transfer function: $T = \dfrac{1}{1 + j \cdot \dfrac{f}{B}}$

If you test this with a source which also has a finite bandwidth, or you cascade the bandwidths in some other manner, you end up with a new transfer function.

$$T = \frac{1}{1 + j \cdot \dfrac{f}{B_1}} \cdot \frac{1}{1 + j \cdot \dfrac{f}{B_2}}$$

Approximate this to a combined equivalent single-pole transfer function.

$$T = \frac{1}{1 + j \cdot \dfrac{f}{B_1}} \cdot \frac{1}{1 + j \cdot \dfrac{f}{B_2}} \approx \frac{1}{1 + j \cdot \dfrac{f}{B}}$$

This approximation can only be applicable in some ‘low frequency’ region, perhaps up to the 3 dB bandwidth. After the second pole, the two-pole response will be rolling off at 40 dB/decade, rather than the 20 dB/decade of the single-pole response.

To look at the equivalence consider the magnitude response and equate the denominators.

$$\sqrt{1 + \left(\frac{f}{B_1}\right)^2} \cdot \sqrt{1 + \left(\frac{f}{B_2}\right)^2} = \sqrt{1 + \left(\frac{f}{B}\right)^2}$$

Square both sides, multiply it out, then subtract 1.

$$\left(\frac{f}{B_1}\right)^2 + \left(\frac{f}{B_2}\right)^2 + \left(\frac{f}{B_1}\right)^2 \cdot \left(\frac{f}{B_2}\right)^2 = \left(\frac{f}{B}\right)^2$$

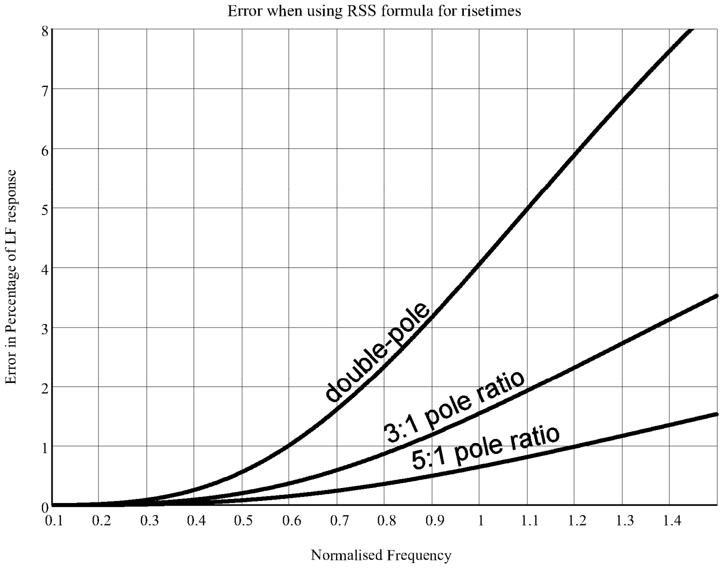
Neglect the cross-product term as being relatively small, then divide by f^2

$$\left(\frac{1}{B_1}\right)^2 + \left(\frac{1}{B_2}\right)^2 = \left(\frac{1}{B}\right)^2$$

Substitute for the risetimes

$$\left(\frac{T_{RB}}{0.35}\right)^2 + \left(\frac{T_{RS}}{0.35}\right)^2 = \left(\frac{T_{RR}}{0.35}\right)^2$$

Giving $T_{RB}^2 + T_{RS}^2 = T_{RR}^2$ then $\boxed{T_{RR} = \sqrt{T_{RB}^2 + T_{RS}^2}}$



These curves show that the errors in using the RSS summing rule become less as the risetimes get more widely separated. This rule is *used too much* and is not useful for situations where the individual responses peak in either the frequency domain or the time domain. The rule says that if you cascade a 200 MHz system and a 400 MHz system, you will get the result:

$$\frac{1}{\sqrt{\dfrac{1}{200^2} + \dfrac{1}{400^2}}} = 179\,\text{MHz}$$

However, if the stages are not simple single-pole responses and are in fact peaked (overshooting) then you can get a resultant bandwidth in excess of 300 MHz.

How are risetime and bandwidth related?

The step response of a single RC circuit is: $V = V_p\left(1 - \exp\left(-\frac{t}{CR}\right)\right)$

Unless otherwise stated, risetime is defined as the time to get from the 10% point to the 90% point on a rising edge. If the waveform has very 'soft' corners, the risetime is sometimes specified from the 20% to 80% points.

For the 10% point: $0.1 = \left(1 - \exp\left(-\frac{t}{CR}\right)\right)$, giving $t_{10\%} = -CR \cdot \ln(0.9)$ and $t_{90\%} = -CR \cdot \ln(0.1)$

$$\boxed{t_{rise} = t_{90\%} - t_{10\%} = CR \cdot \ln(9) = 2.197 \times CR}$$

For a single-pole RC filter the bandwidth $B = \frac{1}{CR}$ in rad/s or $\frac{1}{2\pi CR}$ in Hz.

$$\boxed{bandwidth, B = \frac{1}{2\pi CR} = \frac{2.197}{2\pi \cdot t_{rise}} = \frac{0.350}{t_{rise}}}$$

This formula also works reasonably well for responses that are not from single RC circuits, provided that the pulse responses are 'clean' [minimal undershoot or overshoot].

What is noise bandwidth?

System bandwidth is the band of frequencies over which the response is above –3 dB of the mid-band response. Looking at a single-pole low-pass filter, the frequency response is from DC to the 3 dB bandwidth of *B*. The amplitude response at frequencies above *B* is low, but not zero. There is still a contribution to the system noise from these higher frequencies. For a fairly flat noise spectrum, an effective bandwidth can be ascribed for noise purposes. This new value can then be square rooted and multiplied by the noise density to give the total RMS noise. For the calculation, it is necessary to integrate the magnitude response squared, since it is the sum of squares of noise voltages that is required.

$$B_n = \int_0^\infty \frac{df}{\left|1 + j \cdot \frac{f}{B}\right|^2} = \int_0^\infty \frac{df}{1 + \left(\frac{f}{B}\right)^2} = \left[B \cdot \arctan\left(\frac{f}{B}\right)\right]_0^\infty = B \cdot \arctan(\infty) = \frac{\pi}{2} \cdot B$$

For a simple single-pole system, the noise bandwidth is 1.57× the signal bandwidth. In practice it is unlikely that the system will remain a simple single-pole above 5× or 10× the signal bandwidth, so the figure of 1.5× the signal bandwidth is a more realistic factor.

System	$\frac{\text{Noise Bandwidth}}{\text{3dB Bandwidth}}$
Pure single pole	1.57
Single pole with second pole at 10× first pole	1.44
Double Pole	1.22
Two Pole Butterworth Filter	1.11

Obviously more poles roll the response off faster and the exact shape of the curve governs the multiplication factor.

The points to notice here are:

- the noise bandwidth is *always* greater than the signal bandwidth.
- the difference between noise bandwidth and signal bandwidth becomes less important for a sharp roll-off filter.
- the noise bandwidth is realistically not more than 1.5× the signal bandwidth.

The RMS noise can only be in error by the square root of the noise bandwidth multiplication factor, so the maximum error is typically 20%. This is not a large error when you consider that the noise density is not likely to be that well known or that flat with frequency anyway.

A badly designed system could let the signal response dip to just below the 3 dB limit then continue on at this level for several decades. In this case the noise bandwidth to signal bandwidth ratio would be excessive.

What is the settling time of a resonant circuit?

A resonant circuit will settle to 1% of the initial step in a time given by $\boxed{T_S \approx 2 \cdot Q \cdot T_0}$

Any circuit with the same Q will take the same *number of cycles* of the resonant frequency to settle.

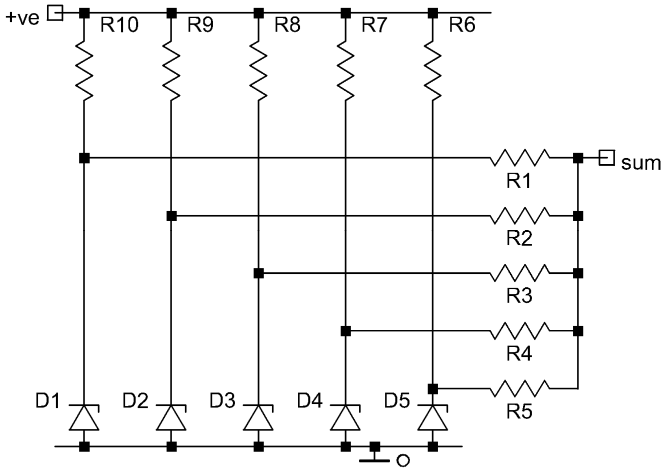
Using the resonant frequency of the tuned circuit $\boxed{T_S \approx \frac{2Q}{f_0}}$

These are very approximate rules and should always be rounded up to the next complete period of the resonant frequency (particularly for a low Q circuit). [For Q>10 the 2 could be replaced by 1.6.] The formula also indicates that simply by counting the number of ***rings*** on the step response, the Q can be estimated.

$\boxed{T_S \approx 10 \cdot \frac{L}{R}}$ series resonant circuit $\boxed{T_S \approx 10 \cdot CR}$ parallel resonant circuit.

How can multiple similar voltages be summed without much error?

A typical application might be to get the average value of several precision references without getting a significant error in the process. This situation has been modelled as a set of zener diodes, the summing being done by nominally equal resistors. The requirement is then to establish the required ratio-matching in the summing resistors in order to get a defined uncertainty in the result. 1 ppm matched resistors should not be used if 1% absolute will do the job!



The voltage summing resistors are R1 to R5. In general there could be *N* of them.

Set the mean summing resistor value as R and the per-unit deviation from the mean as δ. Using similar notation, set the ideal mean input voltage as V and the individual per-unit deviations as ε.

$$V_{IN} \equiv \frac{1}{N} \cdot \sum_{m=1}^{N} V_m = \frac{1}{N} \cdot \sum_{m=1}^{N} \left(V + \varepsilon_m \cdot V\right) = V + \frac{V}{N} \cdot \sum_{m=1}^{N} \varepsilon_m = V$$

… this last step follows from the definition of V as the ideal mean input voltage.

Hence $\sum_{m=1}^{N} \varepsilon_m = 0$, and by the same reasoning $\sum_{m=1}^{N} \delta_m = 0$

Summing currents into the output node (sum) gives:

$$\sum_{m=1}^{N} \frac{V + \varepsilon_m V - V_{SUM}}{R + \delta_m R} = \left(\frac{V}{R} \cdot \sum_{m=1}^{N} \frac{1 + \varepsilon_m}{1 + \delta_m} \right) - \frac{V_{SUM}}{R} \cdot \sum_{m=1}^{N} \frac{1}{1 + \delta_m} = 0$$

$$\therefore V_{SUM} = V\cdot\frac{\sum_{m=1}^{N}\frac{1+\varepsilon_m}{1+\delta_m}}{\sum_{m=1}^{N}\frac{1}{1+\delta_m}} = V\cdot\frac{\sum_{m=1}^{N}(1+\varepsilon_m)\cdot(1-\delta_m+\delta_m^2-\ldots)}{\sum_{m=1}^{N}(1-\delta_m+\delta_m^2-\ldots)}$$

Neglecting third order terms and above:

$$V_{SUM}\approx V\cdot\frac{N-\left(\sum_{m=1}^{N}\delta_m\right)+\left(\sum_{m=1}^{N}\delta_m^2\right)+\left(\sum_{m=1}^{N}\varepsilon_m\right)-\left(\sum_{m=1}^{N}\varepsilon_m\delta_m\right)}{N-\left(\sum_{m=1}^{N}\delta_m\right)+\left(\sum_{m=1}^{N}\delta_m^2\right)}$$

Remembering that the sums of the first order terms are zero:

$$V_{SUM}\approx V\cdot\frac{N+\left(\sum_{m=1}^{N}\delta_m^2\right)-\left(\sum_{m=1}^{N}\varepsilon_m\delta_m\right)}{N+\sum_{m=1}^{N}\delta_m^2} = V\cdot\left(1-\frac{\sum_{m=1}^{N}\varepsilon_m\delta_m}{N+\sum_{m=1}^{N}\delta_m^2}\right)\approx V\cdot\left(1-\frac{1}{N}\cdot\sum_{m=1}^{N}\varepsilon_m\delta_m\right)$$

The per-unit error is : $\boxed{1-\frac{V_{SUM}}{V}=\frac{1}{N}\sum_{m=1}^{N}\varepsilon_m\delta_m}$

Interpreting this answer statistically would give a probable answer, but let's look at the absolute worst case. The sum of all the ε error terms have a mean value of zero, as seen previously; likewise for the δ error terms. The worst possible case would be if half the ε terms had a value of +*e* and the other half had a value of –*e*. Likewise take the worst case of the δ terms as being half +*d* and the other half –*d*. Some of these positive and negative terms would inevitably occur together in the $\varepsilon_m\delta_m$ product, reducing the overall sum. However, being extraordinarily 'unlucky', the ε and δ polarities would align giving

$$\frac{1}{N}\sum_{m=1}^{N}\varepsilon_m\delta_m=\frac{1}{N}\sum_{m=1}^{N}e\cdot d=e\cdot d$$

The absolute worst case for 5% voltage sources and 1% resistors is therefore 0.05×0.01= 500 ppm. Since 3 voltage sources give rise to 6 error terms to be summed, the statistical rules of chapter 3 can be used. You could certainly neglect 20% or more of the error for 3 or more voltage sources. Remember that the error could be due to any source, so the δ term could represent:

- The initial resistor mis-match
- The TC tracking of the resistors
- The long term ratio stability of the resistors

In order to get a summing TC of better than 1 ppm with zeners matched to 5% it would be necessary to use resistors with a TC of 20 ppm/°C or better. 25 ppm/°C resistors may be more readily obtainable and would be quite reasonable due to the statistical improvement of the errors.

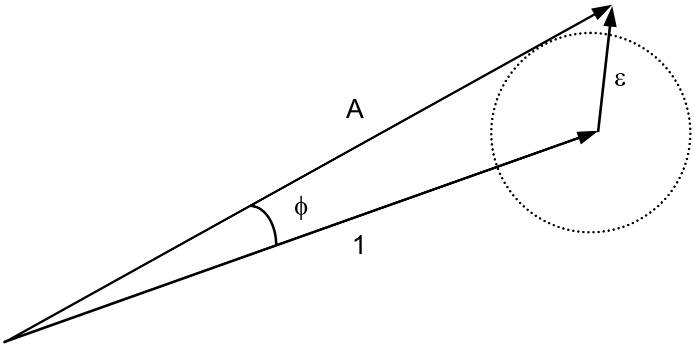
If the long term drift was required to be better than 1 ppm/year the resistors would have to be stable relative to each other to better than 20 ppm/year (for the same 5% mismatch in the voltage sources). Watch out for the drift of the voltage sources however. There is no desensitising effect for drift of the zeners since the average value will still drift.

Why can't the error signal be nulled completely?

Try nulling a sinusoidal signal by subtracting a similar signal from it. The signal won't null completely, regardless of the range and resolution of the trimming pot. One possibility is harmonic distortion on either or both of the signals. The residual error signal can be due to the harmonic content of the waveforms, but this situation will be evident by the error signal being some integer multiple of the fundamental signal frequency. Double the frequency, for example, would mean that the second harmonic distortion was dominant. If the minimum error signal has the same frequency as the input signal then the problem is simply phase shift.

This phasor diagram is exaggerated in order to show the angle and error more clearly.

The signal to be nulled has been normalised to unit amplitude. The nulling signal is of amplitude, *A*, this amplitude being adjusted to minimise the error signal ε. The phase angle ϕ is some fixed amount. By constructing a circle at the end of the signal phasor it is clear that the minimum error signal will occur when the nulling signal and the error signal are at right angles to each other. This result can also be proved with a few lines of calculus.



When the error is optimally nulled, $A = 1 \times \cos(\phi) = \cos(\phi)$. More importantly, $\varepsilon = \sin(\phi)$.

The analysis above also applies to the CMRR of differential amplifiers. The two phasors could represent the inputs to a perfect differential amplifier. Any phase shift would then result in an unwanted signal at the output. The perfect CMRR would have been reduced to $CMRR = 20 \times \log_{10}\left(\frac{1}{\sin(\phi)}\right) = -20 \times \log_{10}(\sin(\phi))$. Remembering that angles in formulae are always expressed in radians, unless otherwise stated, and assuming that the angle is always small,

$$\boxed{CMRR \approx -20 \times \log_{10}(\phi)}.$$

For a single-pole low-pass filter, $T = \frac{1}{1 + j\omega CR}$. This network produces a phase shift $\angle T = -\arctan(\omega CR)$. At frequencies well below the pole, use the small angle approximation $\arctan(\theta) \approx \theta$. Then $\angle T \approx -\omega CR = {}^{f}\!/_{B}$. The (small) phase shift is proportional to frequency.

Thus even when a low-pass filter is 1000× higher than the current operating frequency, it still produces enough phase shift to cause CMRR to be reduced to 60 dB! The CMRR also gets worse at a rate of 20 dB/decade. The amplitude loss in the low-pass filter is of less importance than the phase shift in these calculations.

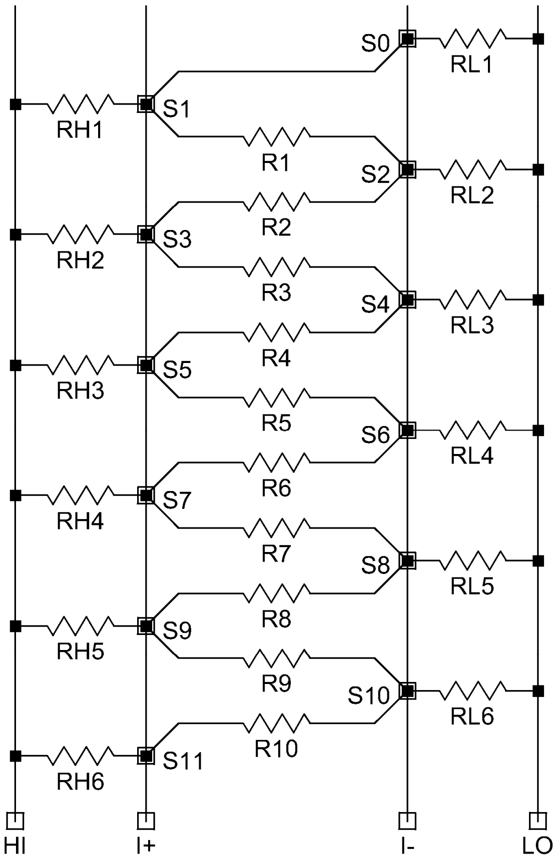
Phase shift also arises due to mismatched path lengths. A one cycle delay is 2π radians of phase shift. Thus a delay of Δt at a frequency *f* (Hz) gives a phase shift $\phi = 2\pi f \times \Delta t$.

$$\boxed{CMRR = -20 \times \log_{10}(2\pi f \times \Delta t)}$$

again getting worse with frequency at a rate of 20 dB/decade.

How does a Hamon transfer standard achieve excellent ratio accuracy?

A Hamon transfer standard[2] is a series chain of matched resistors which can selectively be connected in parallel. This device then achieves a very accurate ratio between the series and parallel values.



One of the key parts of the Hamon transfer standard is the mechanism to selectively connect 4 wires together at a single node, without introducing significant contact resistance. This was originally done with Mercury, but can now be done by some ingenious mechanical design. For now I will just assume the existence of a 4-terminal switch, *S*, with negligible resistance.

This circuit diagram is *unusual* and needs careful explanation. The main resistors; R1, R2, R3 &c, are permanently connected in series; the switches, S1 &c, have a permanent link in the series path between the main resistors.

When operated, the switches S1 &c, join all four wires together at that node. Remember that two of the four wires are already connected together.

Consider the left hand connection of R2; it seems to connect to R3, RH2 and the I+ line. These connections are actually only made when S3 is closed. When S3 is open, R2 is only connected to R3.

The measurement system is being represented by the 4-wire DMM connections HI, I+, I– and LO. Note that the sense resistors RH1, RH6 and RL1 need to be twice the value of the other sense resistors since they are only connected to one main resistor.

To measure all ten resistors in series it is only necessary to close switches S0 and S11. To measure all ten resistors in parallel, all the switches are closed apart from S0. Individual resistors can also be checked to ensure that they are correct. For example R7 would be measured by closing S7 and S8 only.

If the nominal resistance of elements R1 to R10 is R, then the lowest resistance is R/10 and the highest is 10·R. This means that a nominal ratio of 100:1 is achievable. However, there are a large number of other series/parallel combinations so that the total number of different ratios available in the range from 1 to 100 is huge. A computer program is the only sensible way of evaluating the different ratios available.

To look at the accuracy of the transfer ratio, set the mean resistance of the elements R1 to R10 as R. The difference between an individual resistance and the mean can be written using the per-unit error δ. Thus $R_1 = R + \delta_1 R$, $R_2 = R + \delta_2 R$, $R_3 = R + \delta_3 R$, &c.

For the ten resistors in series: $$R_S = \sum_{n=1}^{10} R_n = \sum_{n=1}^{10} \left(R + \delta_n R\right) = 10R + R \cdot \sum_{n=1}^{10} \delta_n = 10R$$

… by the definition of R as the mean, the sum of the deltas is necessarily zero.

[2] B.V. Hamon, 'A 1-100 Ω Build-Up Resistor for the Calibration of Standard Resistors', in *Journal of Scientific Instruments*, 31 (Dec 1954), pp. 450-453.

For the ten resistors in parallel:

$$R_P = \frac{1}{\sum_{n=1}^{10}\frac{1}{R_n}} = \frac{1}{\sum_{n=1}^{10}\frac{1}{R+\delta_n R}} = \frac{R}{\sum_{n=1}^{10}\frac{1}{1+\delta_n}} = \frac{R}{\sum_{n=1}^{10}\left(1-\delta_n+\delta_n^2-\ldots\right)} \approx \frac{R}{10-\sum_{n=1}^{10}\delta_n+\sum_{n=1}^{10}\delta_n^2}$$

Again the sum of deltas is zero giving:

$$R_P \approx \frac{R}{10+\sum_{n=1}^{10}\delta_n^2}$$

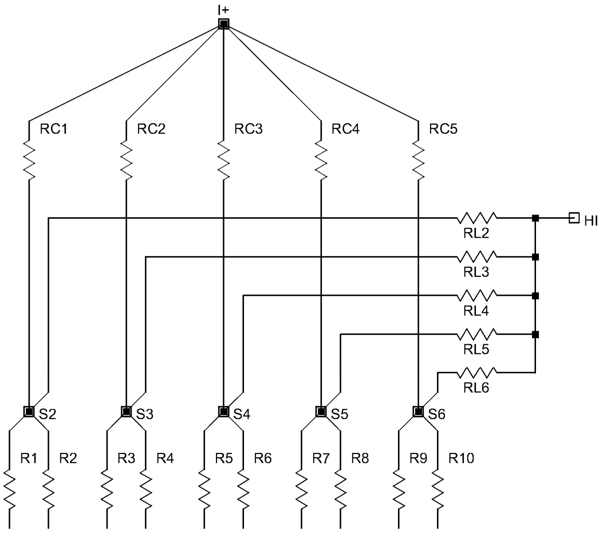
This is a very good approximation when $\delta < 0.01$ since the cubed terms are then more than 100× lower than the squared terms. Since the series chain measures 10R exactly, the parallel formula tells us that if the basic resistors are matched to better than 0.01% (100 ppm), then the 100:1 transfer ratio is better than 0.01 ppm (1 part in 10^8) without even requiring any statistical error reduction.

There are at least four other sources of error to consider:

- The perfection {quality} of the 4-terminal switches.
- The matching of the parallel sense-resistors.
- The matching of the I+ current paths to the switches (this applies to I– current path as well)
- The rejection of the measuring device to changes in the sense resistance.

This last error is easily removed by always ***padding*** the sense resistance up to the maximum value required. If the resistance of one of the sense resistors is RH then the total sense resistance ranges from RH/5 to 2×RH. Hence padding resistance up to 1.8×RH would be needed.

The matching of the parallel sense resistors is intimately connected to the matching of the I+/I– current paths to the switches. This current path would not be physically constructed as shown above because the resistive drop in the I+ conductor would give an offset between resistors RH1 and RH5 in the fully parallel case.



This is a redrawn Hamon transfer standard, just showing the I+ current distribution tree and the HI sense resistors. The current distribution resistors RC1 to RC5 are ideally zero, but the circuit has been specifically drawn to demonstrate that they should be star-connected to the I+ feed point in order to minimise any voltage offsets. If there are significant voltage offsets then the summing resistors RL2 to RL6 need to be better matched.

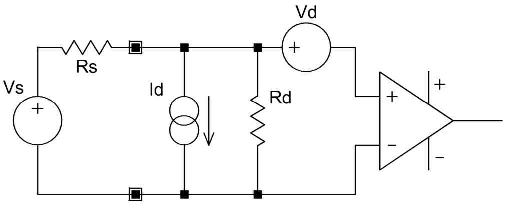
Using the formula for accurately summing voltages from the previous appendix section, if the sense resistors are to be matched to 1% then the per-unit offsets on the current distribution tree need to be of the order of 1 ppm in order to maintain the 0.01 ppm transfer accuracy. In a unit with 100 Ω main resistors, each current feed resistor sees a 50 Ω load; 1 ppm of this is 50 μΩ. Hence this current distribution tree needs to be very carefully matched. The good news is that this matching is very easy to measure and trim, since 1 V across the main resistors gives 1 μV offsets at the switch junctions. This signal level is relatively simple to measure at DC and with low source impedances.

The clever thing about the switches is that they do not need to have zero resistance. If the 4 electrical connections are arranged as a regular tetrahedron, the current path from any point to any other point is matched. Hence the difference in resistance can easily be 10 to 50 times lower than the actual resistance. It is only this difference in resistance that then determines the ratio accuracy.

In order to get the stated transfer accuracy, the individual resistors in the standard have to be run at the same current in each configuration. This stops self-heating errors and voltage coefficient errors in the main resistors compromising the accuracy.

How does impedance matching reduce noise?

Impedance matching means wide-band matching using a transformer, or narrow band matching using some other reactive coupling network. This topic is therefore not relevant to systems which are required to work down to DC.



In this system the signal source V$_s$ has an output resistance R$_s$. The signal detector has an input resistance R$_d$, current noise of I$_d$, and voltage noise of V$_d$. It is obvious that the lowest noise will be obtained when R$_s$ is minimised and R$_d$ is maximised, providing that the detector current and voltage noise sources remain unchanged.

For any particular band of frequencies, a transformer could be placed between the source and the detector. This would have the effect of increasing V$_s$ by the turns ratio, and increasing R$_s$ by the turns ratio squared. Because V$_s$ has been increased, it then makes more sense to consider the relative signal and noise amplitudes. If the signal is doubled, but the total noise is only increased by 20%, the noise has been reduced relative to the signal. It might then be supposed that there was some optimum turns ratio, 1:*n*, which gave the best overall ratio of signal-to-noise.

The voltage at the input to the detector is increased by the transformer turns ratio, then attenuated by R$_d$.

$$signal = nV_S \cdot \frac{R_d}{R_d + n^2 R_S} = \frac{nV_S}{1 + n^2 \cdot \dfrac{R_S}{R_d}} .$$

The noise at the input to the detector is:

$$noise = \sqrt{V_d^2 + \left(\frac{R_d n^2 R_S}{R_d + n^2 R_S} \cdot I_d \right)^2} = \sqrt{V_d^2 + \frac{n^4 R_S^2 I_d^2}{\left(1 + n^2 \cdot \dfrac{R_S}{R_d}\right)^2}}$$

Giving

$$\frac{signal}{noise} = \frac{V_S}{\sqrt{V_d^2 \cdot \dfrac{\left(1 + n^2 \cdot \dfrac{R_S}{R_d}\right)^2}{n^2} + n^2 R_S^2 I_d^2}}$$

Rather than differentiating the whole of the signal-to-noise ratio expression to find the minimum, just look inside the square root sign.

Write:

$$F = V_d^2 \cdot \frac{\left(1 + n^2 \cdot \dfrac{R_S}{R_d}\right)^2}{n^2} + n^2 R_S^2 I_d^2 = V_d^2 \cdot \left(\frac{1}{n^2} + 2\frac{R_S}{R_d} + n^2 \left(\frac{R_S}{R_d}\right)^2 \right) + n^2 R_S^2 I_d^2$$

Then

$$\frac{dF}{dn} = -2\frac{V_d^2}{n^3} + 2nV_d^2 \left(\frac{R_S}{R_d}\right)^2 + 2nR_S^2 I_d^2 ,$$

which is zero when

$$\frac{1}{n^4} = \left(\frac{R_S}{R_d}\right)^2 + \left(\frac{R_S I_d}{V_d}\right)^2$$

giving

$$n^2 = \frac{R_d}{R_S} \cdot \frac{1}{\sqrt{1 + \left(\dfrac{I_d R_d}{V_d}\right)^2}}$$

When $\dfrac{I_d R_d}{V_d} > 3$, $n_{opt} \approx \sqrt{\dfrac{V_d}{I_d R_S}}$. The current noise and voltage noise seen by the detector are therefore equal. $\boxed{\left.\dfrac{signal}{noise}\right|_{\max} \approx \dfrac{V_S}{\sqrt{2V_d R_S I_d}}}$

When $\dfrac{I_d R_d}{V_d} < \dfrac{1}{3}$, $n_{opt} \approx \sqrt{\dfrac{R_d}{R_S}}$. The transformer matching then gives maximum power transfer from the source to the detector. $\boxed{\left.\dfrac{signal}{noise}\right|_{\max} \approx \dfrac{V_S}{2V_d} \cdot \sqrt{\dfrac{R_d}{R_S}}}$

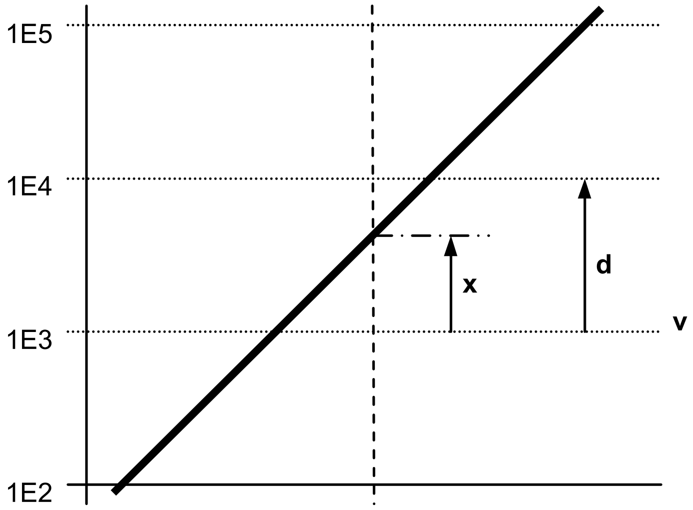
Just make sure you change impedance by using a transformer, or other reactive impedance matching network, rather than by adding resistance. *Adding resistance just increases the noise.*

When dealing with RF and microwave amplifiers, the matching network which gives the *maximum available gain* is probably not the same network that gives minimum noise figure. The curve of minimum noise figure for a microwave amplifier is very misleading for wideband use because it is not possible to make a single matching network cover the entire frequency span. Each point on the minimum noise figure curve is done with a different matching network! Having achieved the minimum noise figure at each frequency, data sheets then plot the gain obtained at each frequency using this optimum noise figure network. The resulting curve is called *associated gain* and it is also an unrealistic representation of the wideband performance of the amplifier.

How do you measure values from a log-scaled graph?

When data sheets only give typical data on log-scaled graphs it is time consuming to work out from first principles how to extract a number from the scale.

Log scales are particularly difficult to read accurately, as they often do not have the minor graduations marked. Even using minor graduations, it is only possible to read the scales to 1 significant figure unless correct interpolation is used.



To plot a value *x* on the graph, you first scale *x* so that it is between 1 and 10 $\{1 \le x < 10\}$. This scaling is done by multiplying or dividing by 10 until *x* is in the desired range. The next step is to take the base 10 logarithm of *x*. Then multiply the result by the size of the single decade graph box that you wish to use.

Suppose for example that the graph has decade boxes of 12 mm. In this case a value of 8 would be plotted as $12 \times \log_{10}(8) = 10.8\,\text{mm}$. This is the distance measured from the “1” box in the direction of the “10” box.

If you now measured the 10.8 (= *x*) it is easy to see that you must divide by 12 (= *d*) and take the inverse-log in order to recover the value of the original point.

As a procedure, first measure the distance, *d*, between the decade boxes. Then measure the distance, *x*, from the nearest *lower* decade box to the point. The value of the lower box, *v* (= 1E3 in the example above), is also required.

$$\boxed{\text{value} = v \times 10^{\left[x/d\right]}}$$

Why is a non-linear device a mixer?

Adding two sinusoidal signals $F_1 = V_1 \cdot \sin(\omega_1 t)$ and $F_2 = V_2 \cdot \sin(\omega_2 t)$, in a linear system does nothing unexpected. Things change when there is a non-linear device in the system.

Consider a transfer function expressed as a power series expansion

$$V_O = A_0 + A_1 \cdot V + A_2 \cdot V^2 + A_3 \cdot V^3 + \ldots$$

When the input is the sum of two sinusoidal input signals, $V = V_1 \cdot \sin(\omega_1 t) + V_2 \cdot \sin(\omega_2 t) = F_1 + F_2$

$$V_O = A_0 + A_1 \cdot [F_1 + F_2] + A_2 \cdot [F_1 + F_2]^2 + A_3 \cdot [F_1 + F_2]^3 + \ldots$$

This is still too large to write down all in one go.

$$V = F_1 + F_2$$

$$V^2 = F_1^2 + 2F_1F_2 + F_2^2$$

$$V^3 = F_1^3 + 3F_1^2F_2 + 3F_1F_2^2 + F_2^3$$

$$V^4 = F_1^4 + 4F_1^3F_2 + 6F_1^2F_2^2 + 4F_1F_2^3 + F_2^4$$

$$V^5 = F_1^5 + 5F_1^4F_2 + 10F_1^3F_2^2 + 10F_1^2F_2^3 + 5F_1F_2^4 + F_2^5$$

These can be immediately written out since they use the binomial coefficients (*Pascal's triangle*).

Consider the V^2 individual terms: $V^2 = F_1^2 + 2F_1F_2 + F_2^2$

$$F_1^2 = V_1^2 \cdot \sin^2(\omega_1 t) = \frac{V_1^2}{2}[1 - \cos(2\omega_1 t)] \qquad \text{double frequency term}$$

$$F_1F_2 = V_1V_2 \cdot \sin(\omega_1 t) \cdot \sin(\omega_2 t) = \frac{V_1V_2}{2}[\cos(\omega_1 t - \omega_2 t) - \cos(\omega_1 t + \omega_2 t)] \qquad \text{difference frequency terms}$$

$$F_2^2 = V_2^2 \cdot \sin^2(\omega_2 t) = \frac{V_2^2}{2}[1 - \cos(2\omega_2 t)] \qquad \text{double frequency term}$$

A (fundamental) mixer ideally produces the sum & difference frequencies $|\omega_1 + \omega_2| \text{and} |\omega_1 - \omega_2|$, the F_1F_2 product being the key term. Thus an ideal ordinary mixer is actually a multiplier. However, a sub-harmonic mixer, as used for mm-waves, may be optimised for $|\omega_{RF} \pm 2 \cdot \omega_{LO}|$, a 3rd order product.

Consider the V^3 individual terms. [$V^3 = F_1^3 + 3F_1^2F_2 + 3F_1F_2^2 + F_2^3$]

$$F_1^3 = V_1^3 \cdot \sin^3(\omega_1 t) = \frac{V_1^3}{4} \cdot [3 \cdot \sin(\omega_1 t) - \sin(3\omega_1 t)]$$

$$F_1^2F_2 = \frac{V_1^2}{2}[1 - \cos(2\omega_1 t)] \cdot V_2 \cdot \sin(\omega_2 t) = \frac{V_1^2V_2}{4}(2 \cdot \sin(\omega_2 t) - \sin(2\omega_1 + \omega_2) - \sin(\omega_2 t - 2\omega_1 t))$$

$$F_1F_2^2 = \frac{V_2^2}{2}[1 - \cos(2\omega_2 t)] \cdot V_1 \cdot \sin(\omega_1 t) = \frac{V_1V_2^2}{4}(2 \cdot \sin(\omega_1 t) - \sin(2\omega_2 + \omega_1) - \sin(\omega_1 t - 2\omega_2 t))$$

$$F_2^3 = V_2^3 \cdot \sin^3(\omega_2 t) = \frac{V_2^3}{4} \cdot [3 \cdot \sin(\omega_2 t) - \sin(3\omega_2 t)]$$

The frequencies generated are $\omega_1, \omega_2, 3\omega_1, 3\omega_2, |2\omega_1 \pm \omega_2|, |2\omega_2 \pm \omega_1|$. The frequencies $3\omega_1, 3\omega_2, |2\omega_1 \pm \omega_2|, |2\omega_2 \pm \omega_1|$, are the third order products. They are of the form $\omega = |n \cdot \omega_1 \pm m \cdot \omega_2|$ where $n + m = 3$, with both *n* and *m* as positive integers.

For modulation, with $\omega_1 >> \omega_2$ the desired frequencies are $\omega_1 \pm \omega_2$. The third order terms $\omega_1 \pm 2\omega_2$ may be within the output bandwidth and may therefore cause problems. They grow as the *second power* of the amplitude of the modulation signal.

For *two-tone intermodulation* tests, $\omega_1 \approx \omega_2$. In this case the third order products $2\omega_1 - \omega_2, 2\omega_2 - \omega_1$ are of special interest since they will be so close to $\omega_1 \text{ and } \omega_2$ that they will be difficult or impossible to filter out. See the picture under ***Intermodulation Distortion***.

Now consider the V^4 individual terms: $V^4 = F_1^4 + 4F_1^3F_2 + 6F_1^2F_2^2 + 4F_1F_2^3 + F_2^4$

$$F_1^4 = \left[F_1^2\right]^2 = \left(\frac{V_1^2}{2}\left[1 - \cos(2\omega_1 t)\right]\right)^2 = \frac{V_1^4}{4}\left[1 - 2\cdot\cos(2\omega_1 t) + \cos^2(2\omega_1 t)\right]$$

$$\therefore F_1^4 = \frac{V_1^4}{4}\left[1 - 2\cdot\cos(2\omega_1 t) + \frac{1 + \cos(4\omega_1 t)}{2}\right] = \frac{V_1^4}{8}\left[3 - 4\cdot\cos(2\omega_1 t) + \cos(4\omega_1 t)\right]$$

$$F_1^3F_2 = \frac{V_1^3}{4}\cdot\left[3\cdot\sin(\omega_1 t) - \sin(3\omega_1 t)\right]\cdot V_2\cdot\sin(\omega_2 t)$$

$$\therefore F_1^3F_2 = \frac{V_1^3\cdot V_2}{8}\left[3\cdot\cos(\omega_1 t - \omega_2 t) - 3\cdot\cos(\omega_1 t + \omega_2 t) - \cos(3\omega_1 t - \omega_2 t) + \cos(3\omega_1 t + \omega_2 t)\right]$$

$$F_1^2F_2^2 = \frac{V_1^2}{2}\left[1 - \cos(2\omega_1 t)\right]\cdot\frac{V_2^2}{2}\left[1 - \cos(2\omega_2 t)\right] \;\therefore F_1^2F_2^2 = \frac{V_1^2V_2^2}{4}\left[1 - \cos(2\omega_1 t) - \cos(2\omega_2 t) + \cos(2\omega_1 t)\cos(2\omega_2 t)\right]$$

$$\therefore F_1^2F_2^2 = \frac{V_1^2V_2^2}{8}\left[2 - 2\cdot\cos(2\omega_1 t) - 2\cdot\cos(2\omega_2 t) + \cos(2\omega_1 t + 2\omega_2 t) + \cos(2\omega_1 t - 2\omega_2 t)\right]$$

Since these equations are symmetric with respect to F$_1$ and F$_2$ just write down the other terms.

$$F_1F_2^3 = \frac{V_1\cdot V_2^3}{8}\left[3\cdot\cos(\omega_1 t - \omega_2 t) - 3\cdot\cos(\omega_1 t + \omega_2 t) - \cos(3\omega_2 t - \omega_1 t) + \cos(3\omega_2 t + \omega_1 t)\right]$$

$$F_2^4 = \frac{V_2^4}{8}\left[3 - 4\cdot\cos(2\omega_2 t) + \cos(4\omega_2 t)\right]$$

The frequencies produced are $|\omega_1 \pm \omega_2|,\ 2\omega_1, 2\omega_2, |3\omega_1 \pm \omega_2|, |3\omega_2 \pm \omega_1|, 4\omega_1, 4\omega_2$

For a modulator, with $\omega_1 >> \omega_2$ the terms $\omega_1 \pm 3\omega_2$ may be in-band and problematic. Notice that these frequencies grow in amplitude with the third power of the modulation amplitude, V_2^3 .

Unfortunately, as far as designing mixers is concerned, we can now have sum and difference frequencies whose amplitude vary according to the third power of the RF voltage ($F_1^3F_2$) or the third power of the LO voltage ($F_1F_2^3$) which make mixer output levels non-linear, and sometimes even non-monotonic, with changes in LO and RF inputs.

Does a smaller signal always produce less harmonic distortion?

Consider a system in terms of the power series expansion of its voltage transfer function:

$$V_O = A_0 + A_1 V_1 + A_2 V_1^2 + A_3 V_1^3 + \ldots$$

For a sinusoidal input signal, $V\cos(\omega t)$, the output is given by:

$$V_O = A_0 + A_1 V \cos(\omega t) + A_2 V^2 \cos^2(\omega t) + A_3 V^3 \cos^3(\omega t) + \ldots$$

The $2^{2n-1}\cos^{2n}(\omega t)$ terms can be expanded in terms of harmonics of the fundamental using standard trig identities. The coefficients of the harmonics can be presented most simply in tabular form. Fractions have been eliminated from the table by multiplying each row by 2^{n-1}, where *n* is the power of the cosine term.

cosine term	HARMONIC AMPLITUDE					
	DC	1	2	3	4	5
$2\cdot\cos^2(\omega t)$	1		1			
$4\cdot\cos^3(\omega t)$		3		1		
$8\cdot\cos^4(\omega t)$	3		4		1	
$16\cdot\cos^5(\omega t)$		10		5		1
$32\cdot\cos^6(\omega t)$	10		15		6	
$64\cdot\cos^7(\omega t)$		35		21		7
$128\cdot\cos^8(\omega t)$	35		56		28	
$2^{2n-2}\cos^{2n-1}(\omega t)$		$\frac{(2n-1)!}{(\text{n-1})!\,n!}$		$\frac{(2n-1)!}{(\text{n-2})!(\text{n}+1)!}$		**
$2^{2n-1}\cos^{2n}(\omega t)$	$\frac{(2n)!}{2\cdot(n!)^2}$		$\frac{(2n)!}{(\text{n-1})!(\text{n}+1)!}$		$\frac{(2n)!}{(\text{n-2})!(\text{n}+2)!}$	

The ** in the table above indicates that the term is not zero, but would not fit. Blank fields are zero.

For any term $\cos^n(\omega t)$, the highest harmonic is the n*th*. Thus for the powers above the 5th, the table does not show all the harmonics. Also notice that the higher order terms produce every other harmonic up to the n*th*. Thus all even-order terms create second harmonic distortion, and all odd order terms create third harmonic distortion.

Consider the coefficients of the second harmonic distortion.

$$V_{O2} = \left(A_2 V^2 \cdot \frac{1}{2} + A_4 V^4 \cdot \frac{4}{8} + A_6 V^6 \cdot \frac{15}{32} + A_8 V^8 \cdot \frac{56}{128} + \ldots \right) \cos(2\omega t)$$

The contribution due to the higher order terms grows at a much greater rate with increasing input voltage. Thus in general, larger signals will tend to give more harmonic distortion. However, the exact values of the coefficients will determine the point at which higher order terms become dominant. Also, if the coefficients have alternating signs, it is quite possible for there to be maxima and minima in the harmonic distortion content. These maxima and minima are readily observed in FFTs of ADC data acquired using pure sinusoidal input signals.

If the input signal is steadily reduced, the second harmonic contribution due to the higher order terms will decrease more rapidly than the contribution from the lower order terms. Ultimately the input-signal-squared term will dominate, but this effect may not be visible if the system noise floor is reached first.

In a spectrum analyser, the power series expansion of the input mixer transfer function gives rise to a simple power series expansion in which the terms all have the same sign and the coefficient of the squared term is larger than the rest. When using a spectrum analyser, therefore, it is essential to check for internally generated distortion by using additional attenuation. Measure the harmonic distortion, add say 2 dB extra attenuation externally, then re-measure the distortion. The fundamental should have dropped by 2 dB. If the second harmonic also drops by 2 dB then the distortion is not due to the spectrum analyser. If the second harmonic drops by 4 dB or more, then the second harmonic was being created by distortion in the mixer of the spectrum analyser. The solution is to use even more input attenuation in order to run the mixer at a lower signal level.

It is not possible to say that for any general signal generator, reducing the signal level will reduce the harmonic distortion. It is true that reducing the signal level in an output stage may reduce the distortion, but if there is a fixed amount of *cross-over distortion* in the output stage then it will become a higher proportion of the output signal as the signal level is reduced. Also, if there are incoming harmonics to the output stage, there could again be cancellation of the harmonics causing the harmonic level generated to be a non-monotonic function of the output voltage.

How much does aperture jitter affect signal-to-noise ratio?

Consider an ideal ADC sampling a pure sinusoidal signal. The quantisation error of the ADC affects the SNR and hence the ENOB. To consider the effects of jitter on SNR and ENOB, it is convenient to neglect the quantisation error in the first instance.

For the purposes of discussion, the incoming sinusoid can be considered as being one cycle long. In practice there might be many cycles acquired, but in that case the sample points could be plotted on this same single cycle, building up a higher and higher density of points as the acquisition time increased. The net result would be a noisy single-cycle waveform, the mean value of which was the pure sine wave.

The time jitter on any particular sampling point translates to an amplitude error, the error magnitude being related to the rate of change of the incoming signal. Specifically, for $V = \sin(2\pi f t)$, a time shift of δt would cause the signal to be in error by

$$\delta V = \sin(2\pi f[t + \delta t]) - \sin(2\pi f t) = 2 \cdot \cos\left(\frac{4\pi f t + 2\pi f \cdot \delta t}{2}\right) \cdot \sin\left(\frac{2\pi f \cdot \delta t}{2}\right).$$

The sampling jitter should be made small compared to the sampling period and it is therefore a good approximation to say that $\delta V \approx 2\pi f \cdot \delta t \cdot \cos(2\pi f t)$ giving

$$(\delta V)^2 = (2\pi f)^2 (\delta t)^2 \cos^2(2\pi f t) = (2\pi f)^2 (\delta t)^2 [1 + \cos(4\pi f t)] \cdot \frac{1}{2}$$

Since the summation is done over a complete cycle of the sampled signal, the cosine term vanishes from the mean-squared result, and hence from the RMS result.

$$\delta V_{RMS} = \delta t_{RMS} \cdot \frac{2\pi f}{\sqrt{2}}$$

It should be noted that the nature of the jitter, whether random or deterministic, does not affect the result.

The peak value of the original sinusoidal signal was 1, making its RMS value $\frac{1}{\sqrt{2}}$. Thus

$$\frac{\delta V_{RMS}}{V_{RMS}} = 2\pi f \cdot \delta t_{RMS}\ , \qquad \boxed{SNR_{dB} = -20 \cdot \log_{10}\left(\frac{\delta V_{RMS}}{V_{RMS}}\right) = -20 \cdot \log_{10}\left(2\pi \cdot f_{signal}\, \delta t_{RMS}\right)}$$

In practice the SNR is also reduced by the quantisation error and the non-linearities of the ADC. If the SNR is measured at some low frequency, say >10,000× lower than that considered for the jitter test, the jitter effect will have been reduced by at least 80 dB, and can therefore be neglected.

$$SNR_{dB} = -10\cdot\log_{10}\left(\frac{\sum\left(\delta V_{RMS}\right)^2}{V_{RMS}^2}\right) = -10\cdot\log_{10}\left(\frac{\left(\delta V_{RMS}\right)^2\Big|_{LF} + \left(\delta V_{RMS}\right)^2\Big|_{JITTER}}{V_{RMS}^2}\right)$$

$$\therefore SNR_{dB} = -10\cdot\log_{10}\left(10^{-\frac{SNR_{LF}}{10}} + 10^{-\frac{SNR_{JITTER}}{10}}\right)$$

In words, you take the LF SNR and the HF SNR and combine them by converting the dB values back to inverse power ratios, adding, and then taking the resulting inverse ratio back to dB form. Suppose the LF SNR is 60 dB and the calculated jitter SNR is 55 dB.

SNR (LF)	=	60 dB → 1×10^{-6}
SNR (JITTER)	=	55 dB → 3.16×10^{-6}
sum		→ 4.16×10^{-6}
SNR (combined)	=	53.8 dB.

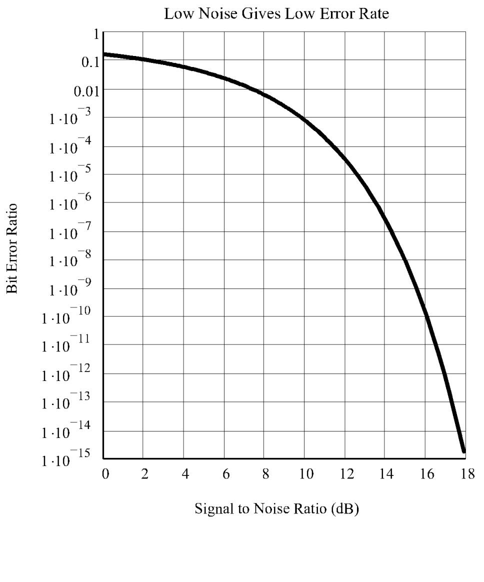
How is Bit Error Ratio related to Signal-to-Noise Ratio?

Bit Error Ratio (BER) is the mean number of errors that occur for a given number of events. It can either be expressed as 1 error in a certain number of events, or it can be given as a per-unit value. Thus a BER of 1 in 10^6 could also be expressed as a BER of 10^{-6}. The synonym *Bit Error Rate* is widespread, but no longer preferred.

Digital communication links are never perfect and a great deal of effort has been spent in creating coding schemes to detect and sometimes to correct these errors. Underlying this is the basic error ratio of the channel, the subject of this section. Even at this hardware level, the method of modulation, the method of detection, and the nature of the channel will affect the exact relationship between BER and SNR. Thus this section is not representative of all systems, but gives an example system to demonstrate the nature of the problem.

The simplest ‘digital’ communication channel is in fact analog, with a detection scheme to decide whether a particular input should be classified as a ‘0’ or a ‘1’. The input data would be modulated in some way on a carrier, transmitted, then demodulated at the other end. For the purposes of this discussion, only the demodulated data will be considered, complete with an amount of added noise.

Suppose the data pattern is deliberately made such that on average the same number of ‘0’s and ‘1’s are transmitted. A filter with a long time constant could then be used to establish the mean level. Signals could then be detected relative to this mean level; signals more positive would be classified as ‘1’s, and signals more negative as ‘0’s. When the signal amplitude relative to the mean level is large, the probability of system noise causing the signal to cross-over the mean level is very low. However, as the signal amplitude becomes progressively lower, it becomes increasingly likely that the noise will cause an incorrect detection of the signal.



The bandwidth presented to the signal and noise will be equal. Consider the signal as a rectangular wave whose amplitude about the mean level is ±*A*. Since the number of ‘1’s and ‘0’s in the signal has been set equal over some long interval, the RMS value of this signal is *A*. If the peak noise exceeds *A* in the one correct direction then an error will be produced. Assuming the noise has a Gaussian (Normal) distribution, the average error rate is found using *cnorm*, the single-tailed cumulative normal distribution function.

Suppose the incoming level is meant to be a logic low at an amplitude of –*A*. In order to get an error the amplitude has to rise to the zero level. The probability of this happening is found by using the normalising expression for the Gaussian distribution and applying the result to the *cnorm* function.

$X = \frac{x-\mu}{\sigma}$, where X is the normalised value, x is the un-normalised value, μ is the mean, and σ is the standard deviation, which in this case is the same as the RMS noise.

$$\text{mean probability of error} = 1 - cnorm\left(\frac{0 - -A}{\sigma}\right) = 1 - cnorm\left(\frac{A}{\sigma}\right)$$

Given that signal to noise ratios are given in terms of power, the above equation can be interpreted as:

$$\boxed{\text{bit error ratio} = 1 - cnorm\left(\sqrt{\frac{\text{signal power}}{\text{noise power}}}\right)}$$

The graph shows that when the signal-to-noise ratio is greater than 11 dB, each dB of SNR improvement gives at least a 10× lower bit error ratio.

What the preceding argument failed to take into account was the time of the sample instant. If the demodulated signal is averaged for any period of time, the noise spikes will be lessened but the signal will be unchanged. The longer the averaging period, the greater the signal-to-noise ratio improvement.

It is useful to look at a theoretical ideal for channel capacity known as the *Shannon-Hartley limit*.

$$\boxed{\text{channel capacity} \le B \cdot \log_2\left(1 + \frac{P_S}{P_N}\right) \text{ bits/second}}$$

B is the channel bandwidth, P_S is the signal power, and P_N is the noise within the channel bandwidth. This limit is for error-free transmission, but in order to approach this limit an arbitrarily complex coding and error correction scheme will be needed.

How are THD, SNR, SINAD and ENOB related?

Total Harmonic Distortion (THD) is a measure of how distorted a sinusoidal signal is.

$$\boxed{\text{THD} = \frac{\text{RSS sum of harmonics}}{\text{fundamental amplitude}}}$$

This is the *ratio* form. Multiply it by 100 to get the THD as a percentage, or take $20\times\log_{10}$ of it to put it in dB form. In an equation, unless otherwise specified, it is this ratio form that is being used.

$$\boxed{THD = \frac{\sqrt{H_2^2 + H_3^2 + H_4^2 + H_5^2 + \ldots}}{H_1} = \frac{\sqrt{\sum_{n=2} H_n^2}}{H_1}}$$

This is what you get when you use a simple notation for the RMS amplitudes of the harmonics with H_n for the n*th* harmonic.

Some manufacturers' definitions of THD only sum the first 5 or 6 harmonic components. This simplification should only be taken as a working approximation and not as a strict definition.

It is possible to substitute the RMS amplitude of the whole waveform, A, for the amplitude of the fundamental. The difference between these two definitions is usually entirely insignificant.

If the THD is δ then $$\frac{A}{H_1} = \frac{\sqrt{H_1^2 + \sum_{n=2} H_n^2}}{H_1} = \sqrt{1+\delta^2}\ .$$

Thus even a 10% THD only gives 0.5% error when making this approximation.

THD does not include the noise on the waveform. In practice, however, one way of measuring THD is to null out the fundamental then measure the RMS value of the remaining signal. In this case the reading does contain the (non-harmonic) noise content as well. This being the case, THD + N (total harmonic distortion plus noise) is defined to take into account the noise that is being measured.

If the RMS value of the noise is denoted by E_{NOISE} then you have:

$$\boxed{THD+N=\frac{\sqrt{E_{NOISE}^2+H_2^2+H_3^2+H_4^2+H_5^2+\ldots}}{H_1}=\frac{\sqrt{E_{NOISE}^2+\sum\limits_{n=2}H_n^2}}{H_1}}$$

If you look at that formula you will see that it is simply the reciprocal of SINAD (the Signal to Noise and Distortion ratio).

$$\boxed{SINAD=\frac{1}{THD+N}}$$

This ratio is often expressed in dB form.

$$SINAD_{dB}=20\cdot\log_{10}\left(\frac{1}{THD+N}\right)=-20\cdot\log_{10}(THD+N)$$

Having a sampled data {digital} version of a distorted sinusoid, it is possible to mathematically do a least-square sine fit and calculate the mean-squared error waveform over an integer number of cycles. In this case:

$$SINAD_{\mathrm{dB}}=10\cdot\log_{10}\left(\frac{\text{mean squared best fit sinusoid}}{\text{mean squared error per cycle}}\right)$$

It is not immediately obvious that these two definitions of SINAD are equivalent. To see the equivalence, it is necessary to write down the original noisy and distorted waveform in mathematical notation (neglecting any DC component).

$$V(t)\equiv E_N(t)+\sum_{n=1}H_n(t)$$

Then create a pure sinusoidal waveform to subtract from this original waveform, creating a mean-squared error waveform.

$$\text{mean squared error}=\frac{1}{T}\cdot\int_0^T[V(t)-k\cdot\sin(\omega t+\phi)]^2\,dt$$

Expand this, grouping terms of the same frequency

$$\text{mean squared error}=\frac{1}{T}\cdot\int_0^T\left[E_N(t)+\left(\sum_{n=2}H_n(t)\right)+\left(H_1(t)-k\cdot\sin(\omega t+\phi)\right)\right]^2 dt$$

The harmonics of the fundamental are not *correlated* to each other; the integral of the product of one harmonic and another, when taken over an integer number of cycles of the fundamental, is exactly zero. The mean-squared error equation therefore simplifies considerably.

$$\text{mean squared error}=\frac{1}{T}\cdot\int_0^T\left[E_N^2(t)+\left(\sum_{n=2}H_n^2(t)\right)+\left(H_1(t)-k\cdot\sin(\omega t+\phi)\right)^2\right]\cdot dt$$

The integral of the product of the noise and any of the harmonics is also zero because the noise is uncorrelated to the harmonics by definition. The mathematically created sine wave is exactly equal to the fundamental in both amplitude and phase when the mean squared error is minimised.

$$\text{least mean squared error}=\frac{1}{T}\cdot\int_0^T\left[E_N^2(t)+\sum_{n=2}H_n^2(t)\right]\cdot dt$$

Thus the two definitions of SINAD are equivalent.

SNR, the Signal-to-noise ratio, does not take into account the harmonic distortion. By this definition:

$SNR = \dfrac{H_1}{E_{NOISE}}$ or in dB form $SNR_{dB} = 20 \cdot \log_{10}\left(\dfrac{H_1}{E_{NOISE}}\right)$

Giving:

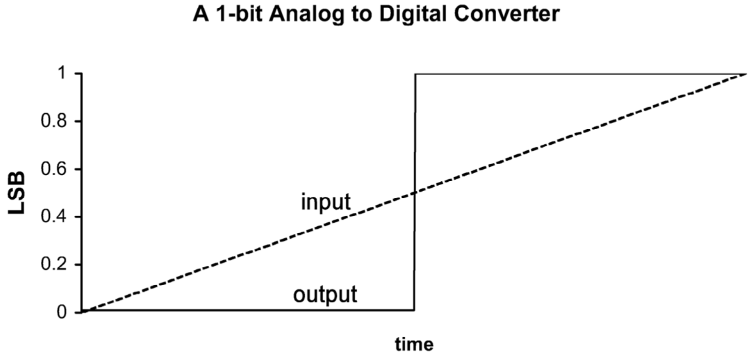
$$\sqrt{\frac{1}{SNR^2} + THD^2} = \sqrt{\frac{E_{NOISE}^2}{H_1^2} + \frac{\sum_{n=2} H_n^2}{H_1^2}} = \frac{\sqrt{E_{NOISE}^2 + \sum_{n=2} H_n^2}}{H_1} = THD + N = \frac{1}{SINAD}$$

or

$$\boxed{SINAD = \frac{1}{\sqrt{\dfrac{1}{SNR^2} + THD^2}}}$$

Do not use this equation with the terms in dB form or percentage form; they must all be done as ratios.

SINAD < SNR, although the difference may not be great if there is a lot more noise than harmonic distortion.



For a perfect 1-bit ADC you can apply a ramp input and you find that for the first half of the ramp the output is low and for the second half the output is high. The conversion system has introduced noise onto the output.

To quantify the noise, calculate the RMS value of the error.

$$NOISE^2 = \frac{1}{1} \cdot \int_0^1 [input(x) - output(x)]^2 \cdot dx = \left(\int_0^{0.5} (x-0)^2 \cdot dx + \int_{0.5}^1 (x-1)^2 \cdot dx\right)$$

$$\therefore NOISE^2 = \left(\left[\frac{x^3}{3}\right]_0^{0.5} + \left[\frac{x^3}{3} - \frac{2 \cdot x^2}{2} + x\right]_{0.5}^1\right) = \left[\frac{1}{24} + \frac{1}{3} - 1 + 1 - \frac{1}{24} + \frac{1}{4} - \frac{1}{2}\right]$$

$$\therefore NOISE = \sqrt{\frac{1}{3} + \frac{1}{4} - \frac{1}{2}} = \sqrt{\frac{4+3-6}{12}} = \frac{1}{\sqrt{12}}$$

For an *N*-bit converter, a full scale sinusoid has an RMS value of $\dfrac{2^N}{2\sqrt{2}}$ LSB.

[The peak-to-peak amplitude is 2^N LSB, the peak is half this and the RMS is $\sqrt{2}$ of the peak.]

The noise has been calculated as $\dfrac{1}{\sqrt{12}}$ LSB / LSB .

This gives the signal-to-noise ratio of an ideal *N*-bit converter as:

$$SNR_{dB} = 20\cdot\log_{10}\left(\frac{\frac{2^N}{2\sqrt{2}}}{\frac{1}{\sqrt{12}}}\right) = 20\cdot\log_{10}\left(2^N\right) + 20\cdot\log_{10}\left(\frac{\sqrt{12}}{2\sqrt{2}}\right) = N\cdot 20\cdot\log_{10}(2) + 20\cdot\log_{10}\left(\sqrt{\frac{3}{2}}\right)$$

at full scale input. $\boxed{SNR_{dB} = 6.0206\times N + 1.7609}$

Smaller signals will have a worse signal-to-noise ratio by an amount of

$$20\cdot\log_{10}\left(\frac{\text{ACTUAL OUTPUT AMPLITUDE}}{\text{FULL - SCALE AMPLITUDE}}\right)$$

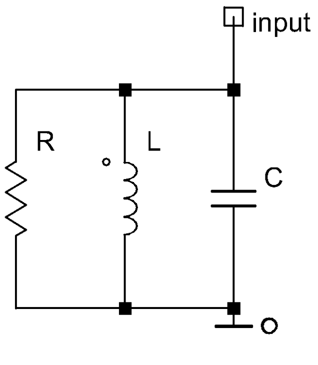
To see how a real converter performs relative to a perfect converter, measure its SINAD and use this in place of the SNR in the above formula. Rearranging gives:

$$\boxed{ENOB = \frac{1}{6.021}\cdot\left[\left(SINAD_{dB} - 1.761\right) + 20\cdot\log_{10}\left(\frac{\text{FULL - SCALE AMPLITUDE}}{\text{ACTUAL OUTPUT AMPLITUDE}}\right)\right]}$$

What is Q and why is it useful?

The Q-factor is the ‘quality’ factor of a resonant component or circuit. It is directly applicable to LCR circuits (the R is either parasitic or is a source/load/bias impedance). The most fundamental definition of Q is: $\boxed{Q = 2\cdot\pi\cdot\frac{\text{peak energy stored}}{\text{energy dissipated per cycle}}}$ (at resonance)

For a parallel LCR circuit:



At resonance the peak energy stored in the inductor is equal to that stored in the capacitor. The peak energy stored is easily calculated as $\frac{1}{2}\cdot C\hat{V}^2$ and the energy dissipated per cycle is simply the product of the mean power and the time of one cycle.

$$Q = 2\pi\cdot\frac{\frac{1}{2}\cdot C\cdot(V\sqrt{2})^2}{\frac{V^2}{R}\cdot\frac{1}{f_0}}$$

$$\boxed{Q_P = 2\pi\, f_0 CR = \omega_0 CR}$$

The 0 subscript simply means at resonance. This formula is often given in terms of the inductance because the inductor, as a component, generally has a lower Q than the capacitor. The peak energy stored in an inductor is $\frac{1}{2}\cdot L\hat{I}^2$. The current is the voltage divided by the impedance, giving the energy stored as

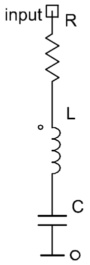
$$\frac{1}{2}\cdot L\hat{I}^2 = \frac{L}{2}\cdot\left(\frac{V\sqrt{2}}{\omega_0\cdot L}\right)^2 = \frac{V^2}{\omega_0^2\cdot L}$$

$$Q = 2\cdot\pi\cdot\frac{\frac{V^2}{\omega_0^2\cdot L}}{\frac{V^2}{R}\cdot\frac{1}{f_0}}$$

$$\boxed{Q_P = \frac{R}{2\pi f_0\cdot L} = \frac{R}{\omega_0 L}}$$

Use this form to convert the Q of the inductor into an equivalent parallel resistance. This equivalent parallel resistance is put in parallel with the load and source resistances to get the overall *loaded* Q of the circuit.

For a series resonant circuit:



$$Q = 2 \cdot \pi \cdot \frac{\frac{1}{2} \cdot L \cdot (I\sqrt{2})^2}{I^2 R \cdot \frac{1}{f_O}}$$

$$\boxed{Q_S = \frac{2\pi f_O L}{R} = \frac{\omega_0 L}{R}}$$

Let’s go back to the parallel circuit. The impedance is:

$$Z_P = \frac{1}{\frac{1}{R} + \frac{1}{j\omega L} + \frac{1}{\left(\frac{1}{j\omega C}\right)}} = \frac{j\omega L}{\frac{j\omega L}{R} + 1 + j\omega C \cdot j\omega L} = \frac{j\omega L}{\left[1 - \omega^2 LC\right] + j\frac{\omega L}{R}}$$

This circuit is resonant when the real term underneath becomes zero, $\omega_0 = \frac{1}{\sqrt{LC}}$.

The impedance formula can be rearranged in terms of the resonant frequency and the Q.

$$Z_P = \frac{j\omega L}{j\frac{\omega L}{R} + \left[1 - \left(\frac{\omega}{\omega_0}\right)^2\right]} = \frac{R}{1 - j\frac{R}{2\pi f \cdot L} \cdot \left[1 - \left(\frac{f}{f_0}\right)^2\right]} = \frac{R}{1 - jQ \cdot \frac{f_0}{f} \cdot \left[1 - \left(\frac{f}{f_0}\right)^2\right]} = \frac{R}{1 - jQ \cdot \left(\frac{f_0}{f} - \frac{f}{f_0}\right)}$$

The 3 dB band edges occur when the term containing the Q becomes ±1. These solutions give the low frequency (f_L) and high frequency (f_H) 3 dB points either side of resonance.

$$Q \cdot \left[\frac{f_0}{f_H} - \frac{f_H}{f_o}\right] = -1 \qquad\qquad Q \cdot \left[\frac{f_0}{f_L} - \frac{f_L}{f_0}\right] = +1$$

We want the bandwidth, $f_H - f_L$. Looking at f_H first:

$$\frac{f_0}{f_H} - \frac{f_H}{f_0} + \frac{1}{Q} = 0 \quad \text{giving} \quad f_H^2 - \frac{f_0}{Q} \cdot f_H - f_0^2 = 0$$

This is now in standard quadratic form, allowing immediate solution:

$$f_H = \frac{+\frac{f_0}{Q} \pm \sqrt{\left(\frac{f_0}{Q}\right)^2 + 4f_0^2}}{2} = f_0 \cdot \sqrt{1 + \frac{1}{4Q^2}} + f_0 \cdot \frac{1}{2Q} = f_0 \cdot \left[\sqrt{1 + \frac{1}{4Q^2}} + \frac{1}{2Q}\right]$$

the solution for f_L is similar

$$\frac{f_0}{f_L} - \frac{f_L}{f_0} - \frac{1}{Q} = 0 \quad \text{giving} \quad f_L^2 + \frac{f_0}{Q} \cdot f_L - f_0^2 = 0$$

$$f_L = \frac{-\frac{f_0}{Q} \pm \sqrt{\left(\frac{f_0}{Q}\right)^2 + 4f_0^2}}{2} = f_0 \cdot \sqrt{1 + \frac{1}{4Q^2}} - f_0 \cdot \frac{1}{2Q} = f_0 \cdot \left[\sqrt{1 + \frac{1}{4Q^2}} - \frac{1}{2Q}\right]$$

To summarise:

$$f_H = f_0 \cdot \left[\sqrt{1+\frac{1}{4Q^2}} + \frac{1}{2Q} \right] \qquad f_L = f_0 \cdot \left[\sqrt{1+\frac{1}{4Q^2}} - \frac{1}{2Q} \right]$$

This gives the bandwidth exactly $\boxed{B = f_H - f_L = \frac{f_0}{Q}}$.

The bandwidth is the centre frequency divided by the loaded Q. A narrow band circuit therefore requires a high Q. The term $\sqrt{1+\frac{1}{4Q^2}}$ is very close to unity for Q >5. Hence the lower and upper frequencies are normally considered to be symmetric about the resonant frequency.

$$f_H - f_0 \approx \frac{f_0}{2Q} \qquad f_0 - f_L \approx \frac{f_0}{2Q}$$

The impedance of a series LCR circuit is:

$$Z_S = R + j\omega L + \frac{1}{jwC} = R \cdot \left[1 + j\frac{\omega L}{R} \cdot \left(1 - \frac{1}{\omega^2 LC} \right) \right],$$

again resonant when $\omega = \frac{1}{\sqrt{LC}}$

giving $$Z_S = R \cdot \left[1 + jQ\frac{f}{f_0} \cdot \left(1 - \left[\frac{f_0}{f} \right]^2 \right) \right] = R \cdot \left[1 + jQ \cdot \left(\frac{f}{f_0} - \frac{f_0}{f} \right) \right]$$

By comparison with the parallel resonant formula, it is clear that the bandwidth rule also applies to the series resonant case.

The definition for Q can be extended by saying that for a component, the reactance is the imaginary part of the impedance, and the resistance is the real part of the impedance.

Whether the reactance is inductive or capacitive: $\boxed{Q = \frac{|\mathrm{Im}\{Z\}|}{\mathrm{Re}\{Z\}}}$,

a convenient form for computer modelling.

What is a decibel (dB)?

The base 10 logarithm of the ratio of two power levels is expressed in Bels, 1 Bel being a 10:1 ratio of powers. This is rather a large unit, so the decibel is preferred. (deci- is the prefix for ×0.1) Named in honour of Alexander Graham Bell who patented the telephone in 1876, the decibel originated around 1923 for long distance telephony.[3] The decibel replaced the “mile of standard cable” that had been used prior to 1923. Initially the name “transmission unit” was used until an international committee settled the issue.[4]

$$\boxed{\text{power ratio in dB} = 10 \cdot \log_{10}\left({}^{P_1}\!/_{P_2} \right)}$$

[3] 'Appendix A: The Decibel and the Neper' in *BR229: Admiralty Handbook of Wireless Telegraphy, Vol 1* (His Majesty's Stationery Office, 1938), pp. 1-2.

[4] W.H. Martin, 'Decibel - The Name for the Transmission Unit', in *The Bell System Technical Journal*, 8 (Jan 1929), pp. 1-2.

If the power is developed across a resistance then $\text{power ratio in dB} = 10 \cdot \log_{10}\left(\frac{V_1^2}{R_1} \cdot \frac{R_2}{V_2^2}\right)$

It is often assumed that R_1 and R_2 are equal, even when they are not, giving a commonly used expression for voltage ratios as:

$$10 \cdot \log_{10}\left(\frac{V_1^2}{R_1} \cdot \frac{R_2}{V_2^2}\right) = 10 \cdot \log_{10}\left(\frac{V_1^2}{V_2^2}\right) = 20 \cdot \log_{10}\left(\frac{V_1}{V_2}\right)$$

To convert a dB figure back into a voltage ratio, divide by 20 and do a base 10 antilog.

Hence 87 dB is a voltage ratio of: $10^{87/20} = 22{,}387$

What is Q transformation?

For a resonant circuit, it may be convenient to view a series impedance as a parallel impedance, or vice versa. At a *single spot frequency* (or over a narrow band of nearby frequencies) a series impedance can be replaced by an equivalent shunt impedance or vice versa.

For a parallel-resonant LC circuit, the magnitude of the inductive and capacitive reactances are equal. The circuit Q can therefore be written as the parallel resistance divided by the reactance, $Q = \frac{R_P}{X_P}$.

Consider a parallel sub-circuit consisting of a resistor R_P and a reactance X_P . The impedance of this combination is:

$$Z = \frac{R_P \times jX_P}{R_P + jX_P} = \frac{R_P \times jX_P}{R_P + jX_P} \cdot \frac{R_P - jX_P}{R_P - jX_P} = \frac{R_P X_P}{R_P^2 + X_P^2} \cdot (X_P + jR_P) = \frac{X_P + jR_P}{{}^{R_P}\!/_{X_P} + {}^{X_P}\!/_{R_P}}$$

$$\therefore Z = \frac{X_P + jR_P}{Q + {}^{1}\!/_{Q}} = \frac{{}^{R_P}\!/_{Q} + jQX_P}{Q + {}^{1}\!/_{Q}} = \frac{R_P}{1+Q^2} + \frac{jX_P}{1 + {}^{1}\!/_{Q^2}} \leftrightarrow R_S + jX_S$$

$$\boxed{\begin{gathered} R_S \cdot \left(1 + Q^2\right) = R_P \\ X_S \cdot \left(1 + \frac{1}{Q^2}\right) = X_P \end{gathered}}$$

These are the series to parallel transformation equations.

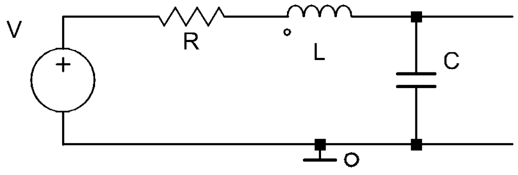
They are *not* approximations and apply even for Q<1.

Why is Q called the magnification factor?

For a series resonant circuit $Q = \frac{\omega_0 L}{R_S}$, with $\omega_0 = \frac{1}{\sqrt{LC}}$.

Combining these equations: $\boxed{Q_S = \frac{1}{\sqrt{LC}} \cdot \frac{L}{R_S} = \frac{1}{R_S}\sqrt{\frac{L}{C}}}$

For a parallel resonant circuit: $\boxed{Q_P = R_P\sqrt{\frac{C}{L}}}$

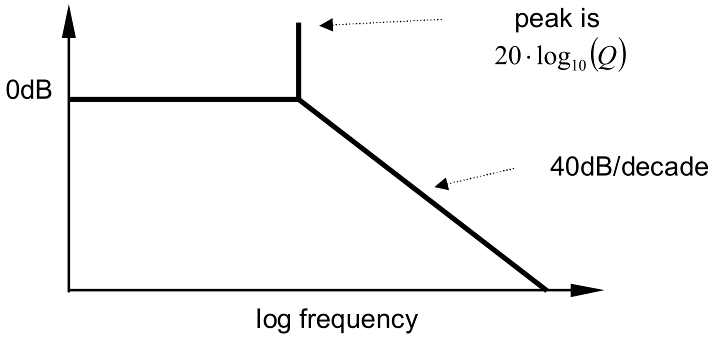


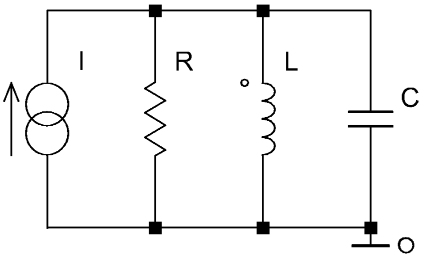
At resonance the impedances of the inductor and the capacitor are equal and opposite. Thus the load on the voltage source is R and the current through the series circuit is $\frac{V}{R}$. The voltage across the capacitor is the current multiplied by its impedance:

$$V_C = \frac{V}{R}\cdot\frac{1}{j\omega_0 C} = -jV\cdot\frac{1}{\omega_0 CR} = -jV\times Q\,.$$

$$\boxed{|V_C| = Q\times|V|}$$

At resonance, the output voltage is therefore the input voltage multiplied by the Q. It should be evident that if the inductor and the capacitor are interchanged the voltage magnification factor is identical. The ***asymptotic*** transfer response of this network is →





For the parallel case, it is the current that is 'magnified'. At resonance the impedances of the capacitor and the inductor cancel so the current source only has a load of R. The voltage across the inductor is therefore $I\times R$.

The current through the inductor is: $I_L = \frac{I\times R}{j\omega_0 L} = -j\cdot I\times Q\,.$ $\boxed{|I_L| = Q\times|I|}$

How can you convert resistive loads to VSWR and vice-versa?

The formula for VSWR contains the modulus of the reflection coefficient. Given that the characteristic impedance is a resistive quantity anyway, the complex numbers reduce to real numbers. The modulus is a positive result achieved by subtracting a smaller quantity from a larger quantity.

For simplicity, write the characteristic impedance as *Z* and the resistive load as *R*, thereby removing the subscripts. When the resistance is larger than the characteristic impedance …

$$VSWR = \frac{1+\left(\frac{R-Z}{R+Z}\right)}{1-\left(\frac{R-Z}{R+Z}\right)} = \frac{R+Z+R-Z}{R+Z-R+Z} = \frac{2\cdot R}{2\cdot Z} = \frac{R_L}{Z_O}$$

$$\boxed{VSWR = \frac{R_L}{Z_0} \text{ for } R_L > Z_0}$$

When the resistance is smaller than the characteristic impedance …

$$VSWR = \frac{1+\left(\frac{Z-R}{Z+R}\right)}{1-\left(\frac{Z-R}{Z+R}\right)} = \frac{Z+R+Z-R}{Z+R-Z+R} = \frac{2\cdot Z}{2\cdot R} = \frac{Z_O}{R_L}$$

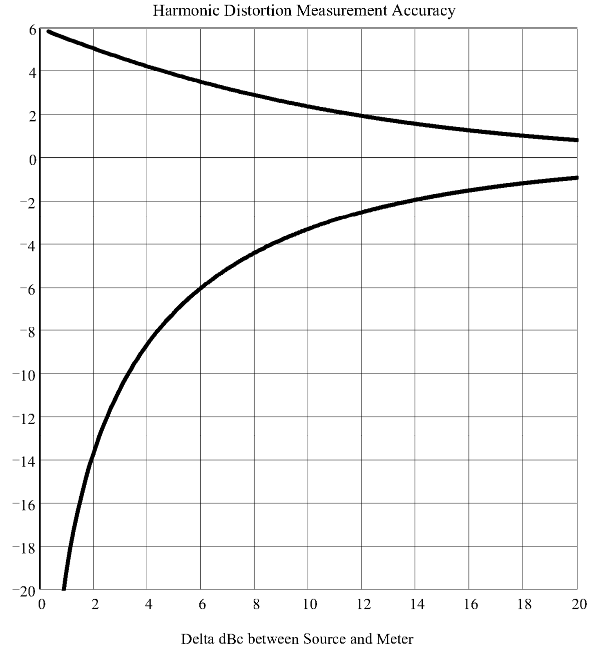
$$\boxed{VSWR = \frac{Z_0}{R_L} \text{ for } R_L < Z_0}$$

What is the Uncertainty in Harmonic Distortion Measurements?

When measuring DC voltage the uncertainty in the measuring instrument is a direct measure of the measurement uncertainty, provided interconnection uncertainties are also considered. A ±10 ppm DMM gives at least ±10 ppm uncertainty in the reading. When considering harmonic distortion the situation is different.

For a harmonic distortion measurement there will be a source and a measuring instrument. For simplicity just consider one harmonic out of the many that may be present. Both the source and the measuring instrument will have some harmonic distortion. Since the harmonics are at a fixed multiple of the fundamental frequency, it is possible to speak of a phase relationship between the harmonic and the fundamental. This could be considered as the phase angle between the rising edge of the fundamental and the rising edge of the harmonic, measured in degrees of the harmonic cycle, for example.

For the source this harmonic would be a real harmonic component. For the measuring instrument considered as a ‘black box’, all that is known is that a pure signal going in produces a spurious harmonic component on its display. Whilst this spurious harmonic might be considered as virtual, it is better to consider it as being a real signal generator with the equipment. It can therefore be nulled by adding an equal amplitude anti-phase harmonic.



The measurement situation is symmetric with respect to the harmonic distortion within the source and measuring devices. If the source has an HD2 of –70 dBc and the meter has an HD2 of –65 dBc, what is the resulting range of the measurement when they are connected together?

Consider the more general situation where the larger harmonic is normalised to unity and is used as the reference phasor. Putting the normalised amplitude of the smaller harmonic as x gives the phasor summation $\sin(n \cdot \omega_C \cdot t) + x \cdot \sin(n \cdot \omega_C \cdot t + \phi)$

The limits for this are simply $\{1+x\}$ and $\{1-x\}$, where x is real and less than unity. x is the normalised harmonic amplitude, which is simply the difference between the two dBc values, expressed as a fraction. The resulting dBc limits for the measurement are therefore:

$$worst = dBc_{worst} + 20 \cdot \log_{10}\left(1 + 10^{-d}\right) \quad \text{and} \quad best = dBc_{worst} + 20 \cdot \log_{10}\left(1 - 10^{-d}\right),$$

where $d \equiv \frac{|\Delta dBc|}{20}$.

For –70 dBc and –65 dBc, $d = 0.25$ and the limits are:

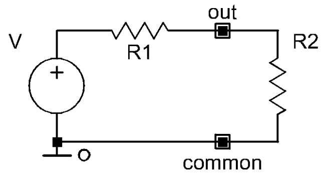
$$worst = -65 + 3.9 = -61.1\,\text{dBc} \qquad best = -65 + -7.2 = -72.2\,\text{dBc}$$

It is instructive to plot the high and low dB error terms as a function of ΔdBc using the worst dBc figure as the nominal. The previous example is read off the scale using 5 dBc on the horizontal axis, yielding limits of +4 dB and –7 dB approximately.

The graph shows that in order to get better than ±3 dB uncertainty in the measured distortion of a source, the meter should be at least 11 dB more accurate. Likewise, in order to test a meter to better than ±3 dB uncertainty, the source has to be at least 11 dB more accurate.

What is RF matching?

For DC circuits, maximum power transfer occurs when load resistance = source resistance.



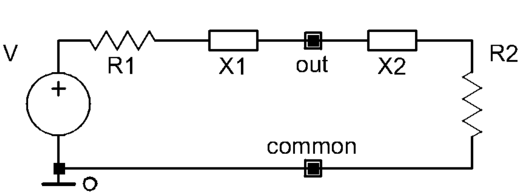
$$P = I^2 R_2 = \left[\frac{V}{R_1 + R_2}\right]^2 \cdot R_2 = \frac{V^2 \cdot R_2}{(R_1 + R_2)^2}$$

Differentiate using the quotient rule, taking R$_1$ and V as constants.

$$\frac{\partial P}{\partial R_2} = \frac{V^2}{(R_1 + R_2)^4} \cdot \left[1 \cdot (R_1 + R_2)^2 - R_2 \cdot 2 \cdot (R_1 + R_2) \cdot (1)\right]$$

$$\therefore \frac{\partial P}{\partial R_2} = \frac{V^2}{(R_1 + R_2)^4} \cdot (R_1 + R_2) \cdot [R_1 + R_2 - 2R_2] = \frac{V^2}{(R_1 + R_2)^3} \cdot (R_1 - R_2)$$

The first derivative is equal to zero only when $R_1 = R_2$. This is seen to be the maximum because setting $R_2 = 0$ or $R_2 = \infty$ both give zero power transfer.



If series reactive elements are present, these need to be equal in magnitude, but opposite in phase, in order to give the maximum current. This would seem to give maximum power transfer when $R_1 = R_2$ as before. To prove it, follow the same route as previously taken.

$$P = I^2 R_2 = \left[\frac{V}{\sqrt{(R_1 + R_2)^2 + (X_1 + X_2)^2}}\right]^2 \cdot R_2 = V^2 \cdot \frac{R_2}{(R_1 + R_2)^2 + (X_1 + X_2)^2}$$

This time there are two variables to adjust; R_2 and X_2.

$$\left.\frac{\partial P}{\partial X_2}\right|_{R_2 \text{ const}} = V^2 \cdot R_2 \cdot \frac{-1}{\left[(R_1 + R_2)^2 + (X_1 + X_2)^2\right]^2} \cdot 2 \cdot (X_1 + X_2)$$

The first derivative of power is zero only when $X_1 = -X_2$. This is a maximum because P=0 when $X_2 = \infty$. Having set $X_1 = -X_2$ the problem reduces to the resistive case; maximum power transfer occurs when $R_1 = R_2$. This is a *conjugate match*; the source impedance is $Z_S = R + jX$ and the load impedance $Z_L = R - jX$, its *complex conjugate*. The notation for a complex conjugate is a superscript *, thus $Z_L = Z_S^*$.

If X_1 is inductive then X_2 is capacitive, and vice versa. Thus the network is AC coupled and there is no power transfer at DC. However, you could do a Q-transformation on the load at a spot frequency, thereby making the AC coupled load into a DC coupled load, should that be desirable.

If R_1, X_1 and X_2 are all fixed, what is the optimum value of R_2 for maximum power transfer?

$$P = V^2 \cdot \frac{R_2}{(R_1 + R_2)^2 + (X_1 + X_2)^2}$$ as before.

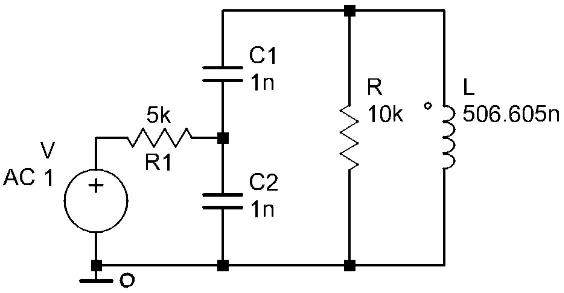
$$\frac{\partial P}{\partial R_2} = \frac{1 \cdot \left[(R_1 + R_2)^2 + (X_1 + X_2)^2\right] - R_2 \cdot 2 \cdot (R_1 + R_2) \cdot 1}{\left[(R_1 + R_2)^2 + (X_1 + X_2)^2\right]^2} = \frac{R_1^2 - R_2^2 + (X_1 + X_2)^2}{\left[(R_1 + R_2)^2 + (X_1 + X_2)^2\right]^2}$$

This is equal to zero when $R_2^2 = R_1^2 + (X_1 + X_2)^2$ giving $R_2 = \sqrt{R_1^2 + (X_1 + X_2)^2}$.

It is a maximum because $R_2 = 0$ gives zero power transfer, as does $R_2 = \infty$.

If you lump X_2 and X_1 together and look at it as the output reactance of the generator, maximum power transfer occurs when the load resistance is equal to the magnitude of the generator's output impedance. This is a useful rule when a conjugate match is not possible.

If you are trying to get a maximised voltage swing across a relatively high impedance, when being driven from a relatively low source impedance, a matching network is needed. Whilst this could be achieved with a transformer, if isolation is not needed an inductive or capacitive network may be adequate, provided the bandwidth requirement is not too great.



For this circuit, R would be the combined loss due to the inductor, the capacitors, and any load across the output. Alternatively, L might be the output itself; the object being to maximise the signal on it.

If L is the output, then C1 and C2 are the matching network. If R is effectively the output, then L is also part of the matching network.

This is a standard matching network, working over a narrow range of frequencies. The analysis is made easier with some substitutions. Putting $X = j\omega_0 L_1$, *X* is then the inductive reactance at resonance.

Resonance occurs when the series reactance of C1 and C2 equals –*X*. Define $k = \frac{C_1}{C_1 + C_2}$.

The reactances of the capacitors are then $X_1 = (k-1)X$ and $X_2 = -kX$.

$$Y_{IN} = \frac{-1}{kX} + \frac{1}{(k-1)X + \frac{RX}{R+X}} = \frac{R+X}{RX + (k-1)RX + (k-1)X^2} - \frac{1}{kX}$$

$$\therefore Y_{IN} = \frac{1}{kX}\left[\frac{R+X}{R+X-X/k} - 1\right] = \frac{1}{kX}\left[\frac{R+X-R-X+X/k}{R+X-X/k}\right] = \frac{1}{kX}\left[\frac{X/k}{R+X-X/k}\right]$$

$$Y_{IN} = \frac{1/k^2}{R+X-X/k} \qquad Z_{IN} = k^2\left(R+X-X/k\right)$$

Remembering that $Q = \frac{R}{\omega_0 L} = j\frac{R}{X}$, $\quad Z_{IN} = k^2\left(R + \frac{jR}{Q}\left(1 - \frac{1}{k}\right)\right) = k^2 R\left(1 - \frac{j}{Q}\cdot\frac{C_2}{C_1}\right)$

For a resonant circuit where the Q is at least several times larger than C_2/C_1 , the input impedance near resonance is: $\boxed{Z_{IN} \approx \left(\frac{C_1}{C_1 + C_2}\right)^2 \cdot R}$, $\qquad L < \frac{R}{3\omega_0}\cdot\frac{C_1}{C_2}$

First set C_1/C_2 to give the desired transformation ratio, calculate *L*, and then finally calculate C_1 & C_2.

What is the difference between a conjugate match and a Z_0 match?

To obtain the maximum power from a source, the *available power*, the load must be the *complex conjugate* of the source impedance. To obtain minimum reflected signal, the load must be matched to the characteristic impedance of the transmission line; this is a non-reflecting or Z_0 matched termination.

The power in the load is $P = I^2 R_L = I^2 \times \mathrm{Re}\{Z_L\}$, where *I* is an RMS value and Re{ } means the real part of the impedance. The current magnitude is evaluated in terms of the magnitude of the voltage source divided by the magnitude of the total impedance in the series circuit.

For microwave/millimetre wave systems it is not possible to directly measure the open-circuit generator voltage and the output impedance. Instead the power delivered to a Z_0 matched (non-reflecting) detector and the output reflection coefficient would be obtained. It is therefore necessary to convert the generator's reflection coefficient into an impedance.

$$\Gamma = \frac{Z_G - Z_0}{Z_G + Z_0} \quad \rightarrow \quad \Gamma(Z_G + Z_0) = Z_G - Z_0 \quad \rightarrow \quad Z_G(\Gamma - 1) = -Z_0(\Gamma + 1) \quad \rightarrow \quad \therefore Z_G = Z_0\left(\frac{1+\Gamma}{1-\Gamma}\right)$$

The power in a Z_0 matched load is $\left|\frac{V_G}{Z_G + Z_0}\right|^2 \times Z_0$, Z_0 being resistive for a lossless line.

The power in the conjugate-matched load is $\left|\frac{V_G}{Z_G + Z_G^*}\right|^2 \times \mathrm{Re}\{Z_G^*\} = \left|\frac{V_G}{Z_G + Z_G^*}\right|^2 \times \frac{Z_G + Z_G^*}{2}$

The ratio of these two powers gives the 'mismatch' loss factor.

$$\eta = \left|\frac{V_G}{Z_G + Z_0}\right|^2 \times Z_0 \times \frac{2\left|Z_G + Z_G^*\right|}{|V_G|^2} = 2Z_0 \times \frac{\left(Z_G + Z_G^*\right)}{\left|Z_G + Z_0\right|^2}$$

But $$Z_G + Z_0 = Z_0\left(\frac{1+\Gamma}{1-\Gamma} + 1\right) = Z_0\left(\frac{1+\Gamma+1-\Gamma}{1-\Gamma}\right) = \frac{2Z_0}{1-\Gamma}$$

And $Z_G = Z_0\left(\frac{1+\Gamma}{1-\Gamma}\right)$; writing $\Gamma = a + jb$

$$Z_G = Z_0 \times \frac{1+a+jb}{1-a-jb} = Z_0 \times \frac{(1+a)+jb}{(1-a)-jb} \cdot \frac{(1-a)+jb}{(1-a)+jb} = Z_0 \times \frac{1-a^2-b^2+j2b}{(1-a)^2+b^2}$$

$$\therefore Z_G + Z_G^* = 2Z_0 \times \frac{1-\left(a^2+b^2\right)}{(1-a)^2+b^2} = 2Z_0 \times \frac{1-|\Gamma|^2}{|1-\Gamma|^2}, \qquad \text{since } |1-\Gamma|^2 = |1-a-jb|^2 = (1-a)^2 + b^2$$

Then finally, $\eta = 2Z_0 \times 2Z_0 \times \frac{1-|\Gamma|^2}{|1-\Gamma|^2} \times \left|\frac{1-\Gamma}{2Z_0}\right|^2$ $\boxed{\therefore \eta = 1-|\Gamma|^2}$

If, instead of being conjugate matched, a source is Z_0-matched, the transferred power is reduced by a factor of $1-|\Gamma|^2$, where Γ is the source reflection coefficient. This is the same factor used for the power reflected from a mismatched load. One application is in antenna theory. The *practical gain* of a transmitting antenna contains the $1-|\Gamma|^2$ factor. The result above shows that the same factor is used for a receiving antenna, hence the practical gain is applicable to both transmitting and receiving antennas.

What is the AC resistance of a cylindrical conductor?

At DC this is a very easy question. The resistance of a cylindrical conductor of diameter *d*, resistivity *ρ*, and length *L* is simply $R_{DC} = \frac{4\rho \cdot L}{\pi \cdot d^2}$. For high frequency alternating currents, where the ***skin depth*** *δ* is only a small fraction of the diameter, the answer is also easy: $R_{HF} \approx \frac{\rho \cdot L}{\pi(d-\delta)\cdot\delta}$. The skin depth is defined in such a way that, to a first approximation, the perimeter of the cross-section times the skin depth gives the equivalent conducting region for any arbitrary shaped conductor

To get the exact resistance of an isolated cylindrical conductor at intermediate frequencies between DC and this high frequency value requires manipulation of Bessel functions with complex arguments, this analysis being first done by Kelvin. The exact formula [5] uses the Kelvin-Bessel functions, ***ber*** and ***bei***, and their first derivatives *ber′* and *bei′*.

$$R_{AC} = R_{DC} \times \frac{x}{2} \cdot \frac{ber(x)bei'(x) - bei(x)ber'(x)}{[ber'(x)]^2 + [bei'(x)]^2}$$

This formula is sufficiently difficult that it is omitted from modern text books. Older texts used to publish the equation results as a table of values.[6] By giving the formula in terms of the parameter *x*, a single table of values covered all cases.

$$x = \pi d \cdot \sqrt{\frac{2f\mu_r}{\rho}} = \frac{d}{\delta\sqrt{2}}$$

For $\frac{d}{\delta} < 10$ the HF approximation is too high by 10% or more.

Butterworth's approximation, $R_{AC} = \frac{R_{DC}}{4} \times \left(1 + \frac{d}{\delta}\right)$, is less than 1.3% too low for $\frac{d}{\delta} > 4$.

For $\left(\frac{d}{\delta}\right) \leq 4.7$:

$$R_{AC} = R_{DC} \cdot \left[1 + \frac{1}{772} \cdot \left(\frac{d}{\delta}\right)^4 \left(1 - \frac{1}{1079} \cdot \left(\frac{d}{\delta}\right)^4 \left[1 - \frac{1}{1687} \cdot \left(\frac{d}{\delta}\right)^4\right]\right)\right]; \qquad < \pm 0.04\% \text{ error.}^7$$

For $\left(\frac{d}{\delta}\right) > 4.7$:

$$R_{AC} = \frac{R_{DC}}{4} \cdot \left[1 + \left(\frac{d}{\delta}\right) + \ln\left[1 + \left(\frac{\delta}{d}\right)\right] \cdot \left(0.77 - 15 \cdot \exp\left[-0.645\left(\frac{d}{\delta}\right)\right]\right)\right]; \qquad < \pm 0.25\% \text{ error.}$$

[5] L.B. Turner, *Wireless: A Treatise on the Theory and Practice of High Frequency Signalling* (Cambridge, 1931), pp. 484-485.

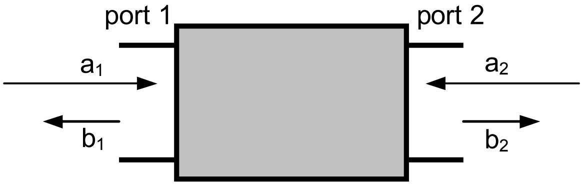
[6] Bureau of Standards, 'Table 17', in *Circular No 74, Radio Instruments and Measurements* (Washington: 1918), p. 309.

[7] L.O. Green, 'Simple Formulae for Skin Effect', in *Electronics World*, 109, no. 1810 (Oct 2003), pp. 44-46.

What are S-parameters?

S stands for *scattering* {separating and driving away in different directions} and has nothing to do with the complex variable **s** found in Laplace transforms. Above say 100 MHz, it is more convenient to look at networks in terms of travelling waves rather than in terms of impedances. Take the example of a fairly long and relatively lossless transmission line terminated in a [wideband] fixed resistor whose resistance is not equal to the characteristic impedance of the line. The magnitude of the input impedance varies cyclically as the frequency increases, but the reflection coefficient stays constant. S-parameters are tightly linked to reflection and transmission coefficients.[8]

For a two-port network there are four S-parameters.



The ports are numbered, rather than labelled *input* and *output*, because the directionality of the device is built into the S-parameters. Usually there would be a common terminal at the input and output [the outer sheath of the coaxial connector for example] but this is not essential.

An *incident wave* arrives at port 1. Some of it is transmitted and some of it is reflected. At the same time, an incident wave is arriving at port 2 and it is also being partially transmitted and reflected. The S-parameters are defined in terms of the following equations:

$$b_1 = S_{11} \times a_1 + S_{12} \times a_2$$

$$b_2 = S_{21} \times a_1 + S_{22} \times a_2$$

As this is a bit complicated to remember, a bit of interpretation is appropriate.[9] These *a* and *b* travelling waves are best considered as voltages in a system with a characteristic impedance of 1 Ω. An alternative viewpoint, used by microwave engineers, is that *a* and *b* are *amplitudes*, normalised such that the amplitude squared equals the power in the travelling wave, $|a| = \sqrt{P}$. Microwave and mm-wave engineers avoid current, voltage and impedance because of the 3-D distribution of the fields, and because the guided wave-impedance varies with frequency (see Waveguide Mode TE_{10}).

If no signal is applied to port 2 $a_2 = 0$, giving $b_1 = S_{11} \times a_1$; b_1 is the reflected signal from an incident signal of a_1. Thus S_{11} is the voltage reflection coefficient when port 2 is not driven and Z_0 terminated. In general, for an N-port system, S_{MM} is the voltage reflection coefficient at port M when all [N–1] other ports are not driven and Z_0 terminated. The ideal value for S_{MM} is zero.

S_{MP} values [$1 \le M \le N$; $1 \le P \le N$] are complex numbers, evaluated in the frequency domain. They are quoted at specific frequencies, or over specific bands of frequency, and can be expected to vary significantly with frequency.

Suppose the two-port network is an amplifier, conventionally having input on port 1 and output on port 2. When you apply a_1 you want to get a larger b_2 as a result. The voltage gain parameter is therefore S_{21}. On the other hand, you do not want the output signal leaking back to the input. Thus the reverse voltage gain factor, S_{12}, is ideally low or zero.

In general, S_{MP} is the complex voltage gain factor *to port M from port P*. If the network is linear and does not contain active devices [amplifiers or rectifiers] ***reciprocity*** means that a particular S-parameter is equal to its transpose; in other words $S_{MP}= S_{PM}$. This reciprocal property should be used with care, however, since passive directional devices such as ***isolators*** are not unusual in the VHF+ region. In any passive system, $|S_{MP}| \le 1$, $|S_{MM}| \le 1$; that is, any S-parameter is less than or equal to unity (conservation of energy).

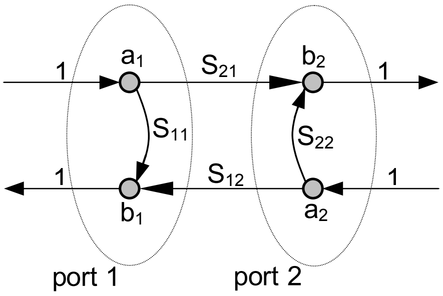
The S-parameter equations can be represented in diagrammatic form as a *signal flow graph*.

[8] H.J. Carlin, 'The Scattering Matrix in Network Theory', in *Institute of Radio Engineers: Transactions on Circuit Theory*, CT-3 (1956), pp. 88-97.

[9] L.O. Green, 'S-Parameters Made Simple', in *Electronics World*, 106, no. 1776 (Dec 2003), pp. 928-933.

$$b_1 = S_{11} \times a_1 + S_{12} \times a_2$$

$$b_2 = S_{21} \times a_1 + S_{22} \times a_2$$



The dotted ellipses and port marking are not normally shown on signal flow graphs. They have been added here to aid understanding. Comparing the equations to the signal flow graph, it is apparent that the nodes have amplitudes. These amplitudes pass to other nodes by branches with defined gain factors.

The incident wave and reflected wave have been separated in space on this diagram, although they still flow down the same conductors (or in the same waveguide). A node vertically above another represents the same physical point in the real world. Note that the S_{11} and S_{22} lines have been shown curved here to help to visualise the reflected nature of these waves.

For a lossless passive two-port network, the mean power going in is equal to the mean power going out. The power balance can be written as … $|a_1^2| + |a_2^2| = |b_1^2| + |b_2^2|$

Driving port 1, with port 2 terminated but not driven, $a_2 = 0$ and $|a_1|^2 = |b_1|^2 + |b_2|^2$

b_1 and b_2 can be expressed in terms of a_1 and a_2, but as a_2 is zero in this case, the power balance equation becomes:

$|a_1|^2 = |S_{11} \times a_1|^2 + |S_{21} \times a_1|^2$ giving $1 = |S_{11}|^2 + |S_{21}|^2$ $\boxed{|S_{21}|^2 = 1 - |S_{11}|^2}$

This result sets a limit on a passive network. $\text{transmitted power} \le \text{incident power - reflected power}$

Since there is no preferred direction implicit in the S-parameter equations, the signals could just as easily pass through the other way. In this case $|S_{12}|^2 = 1 - |S_{22}|^2$

A passive linear two-port network will be reciprocal, making $S_{12} = S_{21}$.

The power balance equations will therefore be equal $|S_{21}|^2 = 1 - |S_{11}|^2 = 1 - |S_{22}|^2 = |S_{12}|^2$

hence $\boxed{|S_{11}| = |S_{22}|}$

The generalisation of the basic equations for 3 or more ports becomes obvious when written in matrix form. The 3 port network equations are:

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

… which can be written in matrix shorthand as $[\mathbf{b}] = [\mathbf{S}][\mathbf{a}]$

For generalised power balance the input power is: $|a_1|^2 + |a_2|^2 + |a_3|^2 + \ldots = [\mathbf{a}]^{T*}[\mathbf{a}]$

… the T* superscript means the transposed {swap rows & columns} complex conjugate matrix.

The output power is: $|b_1|^2 + |b_2|^2 + |b_3|^2 + \ldots = [\mathbf{b}]^{T*}[\mathbf{b}]$

For a loss-less network: $[\mathbf{a}]^{T*}[\mathbf{a}] = [\mathbf{b}]^{T*}[\mathbf{b}] = [\mathbf{a}]^{T*}[\mathbf{S}]^{T*}[\mathbf{S}][\mathbf{a}]$

… from which is derived the *unitary* property of the S-parameters of a lossless network, namely $[\mathbf{S}]^{T*}[\mathbf{S}] = [\mathbf{I}]$, the identity matrix (zero everywhere apart from unity on the leading diagonal).

In general, for a loss-less network, the sum of squared magnitudes of any column in the S-parameter matrix is equal to unity (power in equals power out).

$$1=|S_{11}|^2+|S_{21}|^2+|S_{31}|^2+\dots\ ; \qquad 1=|S_{12}|^2+|S_{22}|^2+|S_{32}|^2+\dots\ ; \qquad 1=|S_{13}|^2+|S_{23}|^2+|S_{33}|^2+\dots$$

The equation $[\mathbf{S}]^{T*}[\mathbf{S}]=[\mathbf{I}]$ also gives a whole load of zero equations. Multiplying them out for a two port network, $S_{11}^*S_{12}+S_{21}^*S_{22}=0$ and $S_{12}^*S_{11}+S_{22}^*S_{21}=0$.

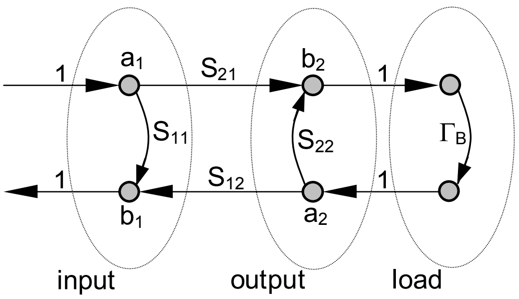
A consequence of all this maths is that it is not possible to create a 3-port lossless matched reciprocal splitter. A linear passive lossless waveguide T-junction can therefore never be fully matched. Matched splitters are therefore made using 4-port structures, with one port terminated. Thus either a ***magic-T*** or a rat-ring {rat-race} can be used as a loss-less matched reciprocal T-splitter.

Why does multiplying two VSWRs together give the resultant VSWR?

Like many ‘rules’, this is an approximation. By knowing what the limits of its use are, you can make a sensible decision about using it.

A typical application is to evaluate the overall VSWR of a load which needs an adaptor fitted to it. This might be to convert from a BNC connector to a type-N connector for example.

The derivation for the relationship between cascaded load VSWRs is explained by this signal flow-graph. The first assumptions to be made are that the adaptor is linear, lossless and reciprocal. From these assumptions you can then state that:



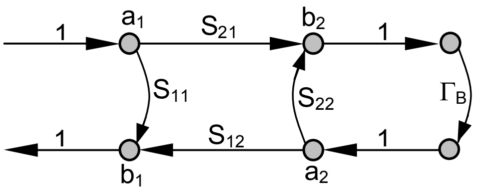
$S_{12}=S_{21}$ (reciprocal network)

$|S_{11}|\equiv|\Gamma_A|$ measured or specified

$|S_{21}|=\sqrt{1-|S_{11}|^2}=\sqrt{1-|\Gamma_A|^2}$ (lossless network)

It is important to realise that $|S_{11}|=|S_{22}|$ (lossless reciprocal network) despite the fact that an adapter cannot be mechanically symmetric from input to output.

Solve the general problem first:



First evaluate b2

$$b_2=S_{21}a_1+S_{22}a_2=S_{21}a_1+S_{22}\Gamma_B b_2$$

$$\therefore b_2=a_1\frac{S_{21}}{1-S_{22}\Gamma_B} \quad\text{and}\quad a_2=a_1\frac{S_{21}\Gamma_B}{1-S_{22}\Gamma_B}$$

$$b_1=S_{11}a_1+S_{12}a_2=S_{11}a_1+S_{12}a_1\frac{S_{21}\Gamma_B}{1-S_{22}\Gamma_B}=a_1\left(S_{11}+\frac{S_{12}S_{21}\Gamma_B}{1-S_{22}\Gamma_B}\right)=a_1\Gamma_R$$

The resultant general input reflection coefficient …

$$\boxed{\Gamma_R=S_{11}+\frac{S_{12}S_{21}\Gamma_B}{1-S_{22}\Gamma_B}}$$

On this specific problem, $|\Gamma_R|\le|\Gamma_A|+|\Gamma_B|\cdot\dfrac{1-|\Gamma_A|^2}{1-|\Gamma_A\Gamma_B|}$

The largest resultant VSWR is therefore:

$$VSWR_R = \frac{1+|\Gamma_R|}{1-|\Gamma_R|} = \frac{1+\left[|\Gamma_A|+|\Gamma_B|\cdot\dfrac{1-|\Gamma_A|^2}{1-|\Gamma_A\Gamma_B|}\right]}{1-\left[|\Gamma_A|+|\Gamma_B|\cdot\dfrac{1-|\Gamma_A|^2}{1-|\Gamma_A\Gamma_B|}\right]} \approx \frac{1+\left(|\Gamma_A|+|\Gamma_B|\right)}{1-\left(|\Gamma_A|+|\Gamma_B|\right)}$$

If you multiply the separate VSWRs you get:

$$VSWR_{AB} = VSWR_A \cdot VSWR_B = \frac{1+|\Gamma_A|}{1-|\Gamma_A|}\cdot\frac{1+|\Gamma_B|}{1-|\Gamma_B|} = \frac{1+(|\Gamma_A|+|\Gamma_B|+|\Gamma_A\Gamma_B|)}{1-(|\Gamma_A|+|\Gamma_B|-|\Gamma_A\Gamma_B|)} \approx \frac{1+\left(|\Gamma_A|+|\Gamma_B|\right)}{1-\left(|\Gamma_A|+|\Gamma_B|\right)}$$

The multiplication of the individual VSWRs gives a result which is fairly close to the correct maximum VSWR provided that both individual VSWRs are less than about 1.2. Above this value the product of the VSWRs is progressively lower than the correct maximum answer. A table of values illustrates the size of the errors.

adaptor VSWR	load VSWR	product of VSWRs	true VSWR
1.05	1.05	1.10	1.10
1.05	1.10	1.16	1.16
1.10	1.10	1.21	1.21
1.10	1.50	1.65	1.66
1.15	1.15	1.32	1.32
1.15	1.20	1.38	1.38
1.20	1.30	1.56	1.57
1.20	1.50	1.80	1.83
1.20	1.80	2.16	2.24
1.25	1.30	1.63	1.64

When measuring the VSWR of inter-series adapters, or such like, it is possible to connect a pair of identical parts back-to-back and then measure the resultant VSWR of the [correctly terminated] pair. The individual VSWRs can then be taken to be the square root of the overall VSWR provided that the overall VSWR is less than about 2.

How are variance and standard deviation related to RMS values?

The RMS value of a voltage is the Root of the Mean of the sum of the Squares:

$$V_{RMS} = \sqrt{\frac{1}{T}\cdot\int_0^T [v(t)]^2\cdot dt}$$

The standard deviation is defined slightly differently.

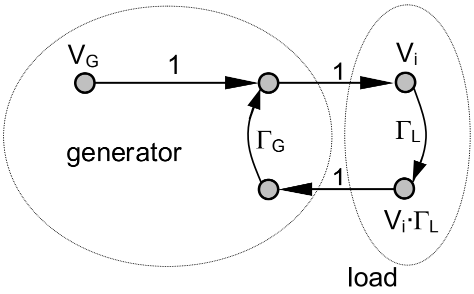
$$\sigma = \sqrt{\frac{1}{T}\cdot\int_0^T \left[v(t)-\overline{v(t)}\right]^2\cdot dt}$$

The standard deviation is the root of the mean of the square of the deviation from the mean. If the mean value of *v(t)* is zero then the RMS value and the standard deviation are the same. Alternatively, standard deviation can be thought of as the AC RMS value. Variance is simply σ^2.

When taking discrete data points, the standard deviation formula is slightly changed from the RMS formula. Instead of dividing by the number of data points, the standard deviation formula divides by one minus the number of data points. For $n > 10$ the difference between the two definitions is only $\frac{50}{n}\%$.

How does an attenuator reduce Mismatch Uncertainty?

Consider the signal flow diagram of a generator feeding a load.



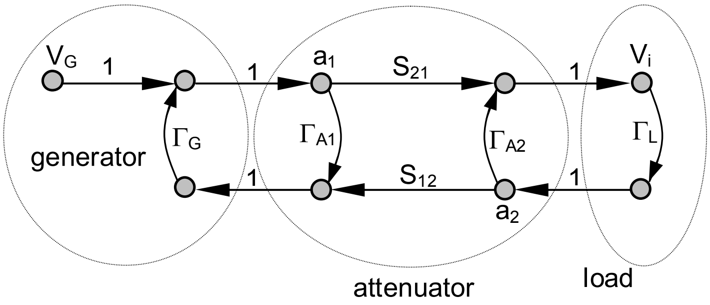
$$V_i = V_G + V_i \cdot \Gamma_L \cdot \Gamma_G$$

$$\therefore V_i = \frac{V_G}{1-\Gamma_L\Gamma_G}$$

The term $1-\Gamma_L\Gamma_G$ is the *mis-match error*.

Any simplifying statements will be left until the end so you can follow the signals more easily.

It is helpful to write $S_{11} \equiv \Gamma_{A1}$ and $S_{22} \equiv \Gamma_{A2}$. This mixed notation emphasises the fact that these S-parameters are just the reflection coefficients of the attenuator when the other end is Z$_0$-matched.



$$a_1 = V_G + a_1\cdot\Gamma_{A1}\cdot\Gamma_G + a_2\cdot S_{12}\cdot\Gamma_G = V_G + a_1\cdot\Gamma_{A1}\cdot\Gamma_G + S_{12}\cdot\Gamma_G\cdot V_i\cdot\Gamma_L \qquad \therefore a_1 = \frac{V_G + V_i\cdot S_{12}\cdot\Gamma_G\Gamma_L}{1-\Gamma_G\Gamma_{A1}}$$

$$V_i = a_1\cdot S_{21} + V_i\cdot\Gamma_L\Gamma_{A2} \qquad \therefore V_i = a_1\cdot\frac{S_{21}}{1-\Gamma_L\Gamma_{A2}}$$

replace a$_1$ $\quad V_i = \dfrac{V_G + V_i\cdot S_{12}\cdot\Gamma_G\Gamma_L}{1-\Gamma_G\Gamma_{A1}}\cdot\dfrac{S_{21}}{1-\Gamma_L\Gamma_{A2}}$, which is a hideous mess to re-arrange

write $V_i = \dfrac{A+B\cdot V_i}{C}$ then $V_i - \dfrac{B}{C}V_i = \dfrac{A}{C}$ and $V_i = \dfrac{\frac{A}{C}}{1-\frac{B}{C}} = \dfrac{A}{C-B}$

$$V_i = \frac{S_{21}\cdot V_G}{(1-\Gamma_L\Gamma_{A2})\cdot(1-\Gamma_G\Gamma_{A1}) - S_{21}\cdot S_{12}\cdot\Gamma_G\Gamma_L} \qquad \boxed{V_i = \frac{S_{21}\cdot V_G}{1-\Gamma_L\Gamma_{A2}-\Gamma_G\Gamma_{A1}-\Gamma_G\Gamma_L\cdot(S_{21}S_{12}-\Gamma_{A1}\Gamma_{A2})}}$$

This equation shows that there is a mismatch at the attenuator input, there is a mismatch at the attenuator output, and the load–generator mismatch is doubly reduced by the attenuation.

For a symmetrical reciprocal attenuator $\Gamma_{A2} = \Gamma_{A1} = \Gamma_A$ and $S_{21} = S_{12}$

$$\boxed{V_i = \frac{S_{21}\cdot V_G}{1-\Gamma_A\cdot(\Gamma_L+\Gamma_G)-\Gamma_G\Gamma_L\cdot(S_{21}^2-\Gamma_A^2)}}$$

For the case without the attenuator, if $|\Gamma_L| = |\Gamma_G| = 0.3$

$|V_i|_{MAX} = \dfrac{|V_G|}{1-|\Gamma_L\Gamma_G|} = 1.099\cdot|V_G|$ and $\quad |V_i|_{MIN} = \dfrac{|V_G|}{1+|\Gamma_L\Gamma_G|} = 0.917\cdot|V_G|$.

There is roughly ±10% uncertainty. Since cable delays result in phase shifts which vary with frequency, the incident voltage can fluctuate over this ±10% band as the frequency changes.

Now consider the case when using a 10 dB attenuator ($\left|S_{21}\right| = 0.316$) having a VSWR= 1.02, ($\left|\Gamma_A\right| = 0.01$). This gives:

$$\left|V_i\right|_{MAX} = \frac{\left|S_{21}\right| \cdot \left|V_G\right|}{1 - \left|\Gamma_A\right| \cdot \left(\left|\Gamma_L\right| + \left|\Gamma_G\right|\right) - \Gamma_G \Gamma_L \cdot \left(\left|S_{21}^2\right| + \left|\Gamma_A^2\right|\right)} = 1.015 \times \left|S_{21}\right| \cdot \left|V_G\right|$$

$$\left|V_i\right|_{MIN} = \frac{\left|S_{21}\right| \cdot \left|V_G\right|}{1 + \left|\Gamma_A\right| \cdot \left(\left|\Gamma_L\right| + \left|\Gamma_G\right|\right) + \Gamma_G \Gamma_L \cdot \left(\left|S_{21}^2\right| + \left|\Gamma_A^2\right|\right)} = 0.985 \times \left|S_{21}\right| \cdot \left|V_G\right|$$

A ±10% mismatch uncertainty has been reduced to a ±1.5% uncertainty, which was the whole point of the exercise. The cost is 10 dB loss of signal.

What is the difference between Insertion Loss and Attenuation Loss?

Using the notation and working from the previous section, the initial incident voltage was:

$$V_i = \frac{V_G}{1 - \Gamma_L \Gamma_G}.$$

Adding the attenuator reduced the incident voltage to $V_i = \dfrac{S_{21} \cdot V_G}{1 - \Gamma_A \cdot \left(\Gamma_L + \Gamma_G\right) - \Gamma_G \Gamma_L \cdot \left(S_{21}^2 - \Gamma_A^2\right)}$

The ratio of these two incident voltages is the numerical value of the insertion loss factor, which is normally expressed in dB. Thus

$$insertion\ loss = 20 \times \log_{10} \left| \frac{V_G}{1 - \Gamma_L \Gamma_G} \times \frac{1 - \Gamma_A \cdot \left(\Gamma_L + \Gamma_G\right) - \Gamma_G \Gamma_L \cdot \left(S_{21}^2 - \Gamma_A^2\right)}{S_{21} \cdot V_G} \right|$$

Note the magnitude brackets inside the log term. You don't want to take the log of a complex number.

$$\therefore insertion\ loss = -20 \times \log_{10} \left| S_{21} \cdot \frac{\left(1 - \Gamma_L \Gamma_G\right)}{1 - \Gamma_A \cdot \left(\Gamma_L + \Gamma_G\right) - \Gamma_G \Gamma_L \cdot \left(S_{21}^2 - \Gamma_A^2\right)} \right|$$

The insertion loss is therefore not only a function of the attenuator, it is also a function of the mismatch at either end. If you want to specify only the attenuator then you can set $\Gamma_L = \Gamma_G = 0$. It should be clear that the complicated insertion loss formula reduces down to a very simple attenuation loss formula.

$$\boxed{attenuation\ loss = -20 \times \log_{10} \left|S_{21}\right|}$$

In this way, an attenuator can be calibrated in terms of attenuation loss, $|S_{21}|$, and reflection coefficients, S_{11} & S_{22}. When used in a real situation, the multiple mismatch uncertainty needs to be calculated to give the actual insertion loss. In practice it is likely that only magnitudes of S_{21}, S_{11}, S_{22}, Γ_G and Γ_L would be known. Thus it would be necessary to work out the actual insertion loss as a band of possible values.

Does a short length of coax cable always attenuate the signal?

For RF interconnections in coax using 50 Ω input and output impedances, cables always attenuate the signal. When operating below 10 MHz with high impedance equipment, however, it is possible for the signal to be larger at the receiving end than at the sending end. Such errors are important when connecting DVMs with calibrators at accuracies better than 5%.

For sinusoids, with a forward wave V_F and a reverse wave V_R, the voltage on the line a distance $-X$ from the load is given by: $V = V_F \exp\left(+\gamma X\right) + V_R \exp\left(-\gamma X\right)$

where the propagation constant $\gamma = \alpha + j\beta$.

A short cable driving a high impedance load can be considered as almost lossless, making $\gamma = j\beta$. The resulting voltage transfer function T is evaluated using two distances, X=0 and X=x.

$$T = \frac{V_{OUT}}{V_{IN}} = \frac{V_F \exp(0) + V_R \exp(0)}{V_F \exp(+j\beta x) + V_R \exp(-j\beta x)}$$

The load Z_L and the characteristic impedance of the line Z_0 give the reflection coefficient Γ, and hence set the ratio of the forward and reverse waves.

$\Gamma = \frac{V_R}{V_F} = \frac{Z_L - Z_0}{Z_L + Z_0}$, which when substituted above simplifies to

$$T = \frac{2Z_L}{(Z_L + Z_0)\exp(+j\beta x) + (Z_L - Z_0)\exp(-j\beta x)}$$

With a purely capacitive load C, at a frequency f, this further simplifies to:

$$T = \frac{1}{\cos(\beta x) - 2\pi f C Z_0 \sin(\beta x)}$$

This is the exact equation for the peaking, but since the line will only be operated up to a small fraction of the resonant peak, the per-unit error can be approximated as:

$$\Delta = T - 1 \approx 1 - \frac{1}{T} = 1 - \cos(\beta x) + 2\pi f C Z_0 \sin(\beta x) \approx \frac{(\beta x)^2}{2} + 2\pi f C Z_0 \beta x$$

The phase constant and physical line length are then replaced by the operating frequency and the propagation delay, t_d, down the cable: $\beta x = 2\pi f t_d$, giving

for Δ<0.05 $$\boxed{\Delta \approx (2\pi f)^2 t_d \left(C Z_0 + \frac{t_d}{2} \right)}$$

How do you establish an absolute antenna calibration?

The key to antenna calibration is the *Friis transmission formula*.[10]

$$\boxed{P_R = P_T \cdot \frac{A_{EA} \cdot A_{EB}}{\lambda^2 r^2} = P_T \cdot G_A \cdot G_B \cdot \left(\frac{\lambda}{4\pi r} \right)^2}$$

where P_R and P_T are the received and transmitted powers respectively. λ is the free space wavelength of the transmission. r is the distance between the antennas, or more accurately the distance between the *phase centres* of the antennas. A_{EA} and A_{EB} are the effective ***apertures*** of the transmit and receive antennas. G_A and G_B are the gains of the transmit and receive antennas relative to a lossless isotropic antenna.

For this formula to work correctly the two antennas need to be aligned for maximum signal strength, including polarisation. The formula assumes a free-space path, without reflections. It also assumes that the antennas are widely separated, $r > \frac{2d}{\lambda}$, d being the maximum linear dimension of either antenna.

The biggest uncertainty in the calibration method can be the choice of *free-space range*. The calibration should ideally be done in a large RF anechoic chamber to simulate free-space conditions and to prevent unauthorised {illegal} radio frequency transmissions. Mounting both antennas on tall masts minimises the ground reflections, but does not eliminate the interference from broadcast transmitters, and does not prevent illegal transmissions.

Using the Friis transmission formula three antennas can be calibrated as a group, the *three antenna*

[10] H.T. Friis, 'A Note on a Simple Transmission Formula', in *Proceedings of the IRE & Waves and Electrons* (May 1946), pp. 254-256.

method. Transmit/receive power ratios are measured for each of the three combinations of antennas, and these power ratios are used to compute the individual gains of the antennas by solving the simultaneous equations. Specifically:

$$\frac{P_{R1}}{P_{T1}} = G_A \cdot G_B \cdot \left(\frac{\lambda}{4\pi r}\right)^2 ; \qquad \frac{P_{R2}}{P_{T2}} = G_B \cdot G_C \cdot \left(\frac{\lambda}{4\pi r}\right)^2 ; \qquad \frac{P_{R3}}{P_{T3}} = G_C \cdot G_A \cdot \left(\frac{\lambda}{4\pi r}\right)^2$$

eliminating G_C from the last two equations gives $G_B = \frac{P_{R2}}{P_{T2}}\left(\frac{4\pi r}{\lambda}\right)^2 \frac{P_{T3}}{P_{R3}}\left(\frac{\lambda}{4\pi r}\right)^2 G_A$

which can now be substituted into the first equation

$$\boxed{G_A = \frac{4\pi r}{\lambda}\sqrt{\frac{P_{R1}}{P_{T1}} \cdot \frac{P_{T2}}{P_{R2}} \cdot \frac{P_{R3}}{P_{T3}}}}$$

The matching of both antennas to their respective connection points is essential. This can be demonstrated by expanding the Friis transmission formula to include the mismatches at the generator, at the antennas, and at the receiver.

$$\boxed{P_R = P_S \frac{\left(1-|\Gamma_S|^2\right)\left(1-|\Gamma_{TA}|^2\right)}{\left(1-|\Gamma_S|\cdot|\Gamma_{TA}|\right)^2} G_{TA} \times \frac{\left(1-|\Gamma_R|^2\right)\left(1-|\Gamma_{RA}|^2\right)}{\left(1-|\Gamma_R|\cdot|\Gamma_{RA}|\right)^2} G_{RA} \times \left(\frac{\lambda}{4\pi r}\right)^2}$$

… where the subscripts are for the source (S), receiver (R), transmitting antenna (TA) and receiving antenna (RA). The four bracketed factors containing only one reflection coefficient variable can be combined with the relevant component as a calibration constant. For example $\left(1-|\Gamma_{TA}|^2\right)\times G_{TA} = G_{PTA}$, where the gain of the transmitting antenna becomes the *practical gain* of the transmitting antenna. The recalibration of the source and received powers has been denoted by appending a ‘C’ subscript.

$$P_{RC} = P_{SC} \frac{G_{PTA}}{\left(1-|\Gamma_S|\cdot|\Gamma_{TA}|\right)^2} \times \frac{G_{PRA}}{\left(1-|\Gamma_R|\cdot|\Gamma_{RA}|\right)^2} \times \left(\frac{\lambda}{4\pi r}\right)^2$$

It should now be clear that interchanging the roles of the transmit and receive antennas will have a negligible effect on the received power, provided the mismatch product terms $|\Gamma_S|\cdot|\Gamma_{TA}|$, $|\Gamma_\mathrm{R}|\cdot|\Gamma_{RA}|$, $|\Gamma_\mathrm{S}|\cdot|\Gamma_{RA}|$, $|\Gamma_\mathrm{R}|\cdot|\Gamma_{TA}|$ are all negligible. For a test system this condition is achieved by using low reflection coefficient attenuators at the output of the source and at the input of the receiver, thereby Z_0-matching both the source and receiver. Since the results should then be identical with each antenna used as either the transmit or receive antenna, it is better to make measurements of all 6 permutations of antenna position, averaging the results for each pair of antennas.

The received power is measured using a well matched power meter. Having measured the received power, the power meter is then moved up to transmitter. The transmit power level is measured by uncoupling the transmitting antenna and feeding the transmitter through an attenuator into the power meter. The attenuator is adjusted to give a similar reading on the power meter to that achieved in the receive position. This technique gives the most accurate ratio of transmit and receive power levels.

In practice the three antenna method would be used by a calibration laboratory to establish a standard against which other antennas could be compared by substitution. This substitution calibration method is done by simply setting up an RF link using the standard antenna for transmission and some other antenna for the reception. Swapping the standard antenna for the antenna to be calibrated gives a very simple gain calibration in terms of the ratio of the received powers and the gain of the standard antenna.

Alternatively two nominally identical antennas can be calibrated as a pair, using one application of the Friis transmission formula. In this case it is still wise to finally compare them both to some other antenna to quantify just how ‘identical’ they are.

How is the coupling coefficient calculated for a tapped inductor?

In a tapped inductor it is certain that the mutual inductance aids the overall inductance. Use the formula $L_{\mathrm{TOTAL}} = L_1 + L_2 + 2M$. Calculate L_1, L_2 and L_{TOTAL} directly from Nagaoka's formula and then use $M = \frac{L_{\mathrm{TOTAL}} - L_1 - L_2}{2}$. Given that the coupling coefficient is defined as $k = \frac{M}{\sqrt{L_1 L_2}}$ the coupling coefficient can then be written as $k = \frac{M}{\sqrt{L_1 L_2}} = \frac{L_{TOTAL} - L_1 - L_2}{2\sqrt{L_1 L_2}}$. Thus the coupling coefficient can be calculated using self-inductance formulae.

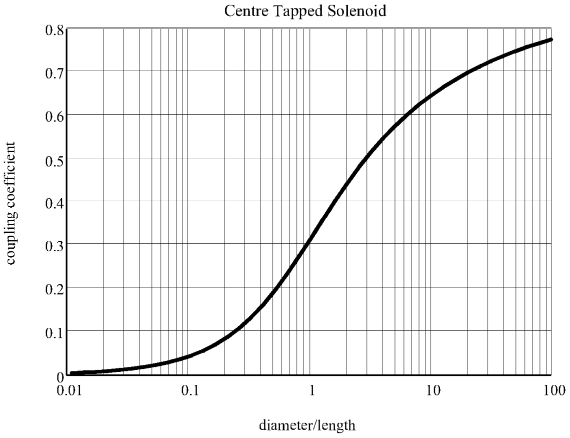
The special case of a centre-tapped single layer air-cored inductor is of interest for peaking circuits. In this case it is desirable to create a tapped inductor with a given coupling coefficient.

For a centre tapped coil $L_1 = L_2$, so $k = \frac{L_{TOTAL}}{2 \cdot L_1} - 1$.

Whilst air-cored inductors are not as popular for modern radio receivers as they used to be, they are easy to calculate and can be used to test-out designs without tooling up expensive prototypes. They also do not suffer from non-linearity and intermodulation products.

For air-cored inductors the coupling is taken into account by the factor *F* in the equation $L = F \cdot n^2 d$. Substitute this for the full and half coils that constitute the tapped inductor, realising that the half coil has half the number of turns of the full coil.

$$\boxed{k = \frac{L_{TOTAL}}{2 \cdot L_1} - 1 = \frac{F_{FULL} \cdot (2n)^2 d}{2 \cdot F_{HALF} \cdot n^2 d} - 1 = 2 \cdot \frac{F_{FULL}}{F_{HALF}} - 1}$$



The value of *k* can then be calculated or plotted. The graph shows that, as expected, very short coils have the best coupling. For a pair of isolated air-cored coils it is difficult to get a coupling coefficient above 0.7. In fact a coupling coefficient of greater than 0.4 would be considered as "close coupling". This is a fairly loose definition! Loose coupling might be defined as *k* less than 0.01, but again this definition is subject to great personal interpretation.

Why is charging a capacitor said to be only 50% efficient?

This was a 'traditional' view, dating back over 100 years, *which is not correct for typical usage*. This rule is only true when the capacitor is charged from zero using a voltage source. It is assumed that there is a pure resistance in the charging circuit, although this resistance can be made arbitrarily small. The integral of the current is the final charge on the capacitor.

$$\text{Energy input} = \int_0^T v(t) \cdot i(t) \cdot dt = V \cdot \int_0^T i(t) \cdot dt = V \cdot q = C \cdot V^2 \qquad \text{Energy stored in capacitor} = \frac{1}{2} \cdot C \cdot V^2$$

$$\boxed{\therefore \text{Charging Efficiency} = 100\% \cdot \frac{\frac{1}{2} \cdot C \cdot V^2}{C \cdot V^2} = 50\,\%}$$

Charging via an inductor, from a current source, or from a sequence of voltage sources each say 20% larger than the previous one, all dramatically improve the efficiency. Work through EX 6.10.1 to get an example of higher efficiencies.

What coax impedance is best for transferring maximum RF power?

For maximum RF power transfer the characteristic impedance of the cable will be equal to the load resistance, $Z_O = R_L$ [a matched load means no reflected power]

The power transferred to the load is: $P = \frac{V^2}{R_L} = \frac{V^2}{Z_0}$ [provided the coax is lossless]

From this formula it would appear that R_L needs to be minimised in order to maximise the power transfer, but that does not take into account the characteristic impedance of the coaxial cable. The characteristic impedance of a coaxial cable is dependant on three factors:

i. the outer radius of the inner conductor, r_I

ii. the inner radius of the outer conductor, r_O

iii. the dielectric constant of the insulator, ε_r

$$\boxed{Z_O = \frac{1}{2\pi} \cdot \sqrt{\frac{\mu}{\varepsilon}} \cdot \ln\left(\frac{r_O}{r_I}\right) \approx \frac{60}{\sqrt{\varepsilon_r}} \cdot \ln\left(\frac{r_O}{r_I}\right)}$$

For a given dielectric and size of cable, the only available parameter to adjust is the outer radius of the inner conductor, r_I. Z_0 is made lower by making r_I larger. What has not so far been considered, is the voltage that can be applied to the coax. Regardless of the dielectric used, the greatest voltage stress on it comes at the interface to the inner conductor. All the electric flux is concentrated at this point.

The electric field intensity is inversely proportional to the radius, $E = \frac{dV}{dr} = \frac{k}{r}$

Integrate from r_I to r_O to get the applied voltage:

$$\int_0^V dV = k \cdot \int_{r_I}^{r_O} \frac{dr}{r} \qquad \text{giving} \qquad V = k \cdot \ln\left(\frac{r_O}{r_I}\right)$$

The *k* from this equation can now be substituted back into the electric field equation, giving the maximum voltage stress as:

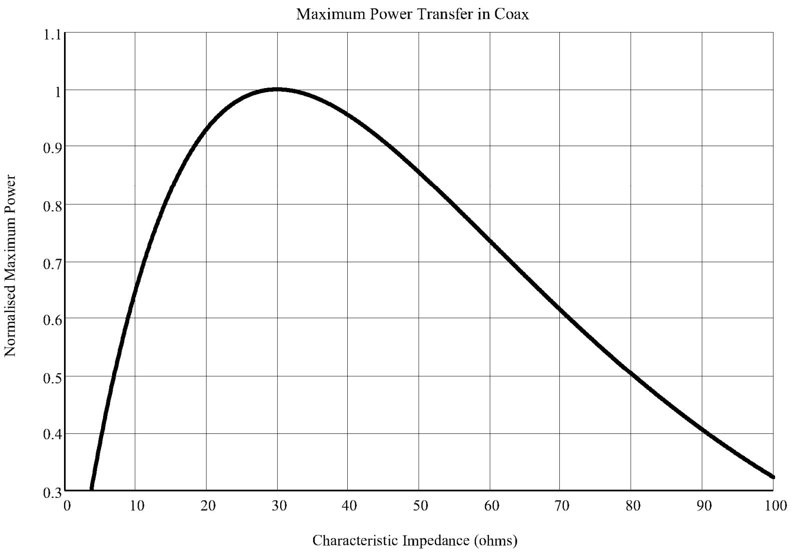
$$E_{MAX} = \left.\frac{dV}{dr}\right|_{MAX} = \frac{k}{r_I} = \frac{V_{MAX}}{r_I \cdot \ln\left(\frac{r_O}{r_I}\right)}$$

Recognise that this is a peak voltage and you will actually be working in terms of the RMS voltage of the applied RF sinusoidal signal. Combining these results:

$$\hat{P}_{MEAN} = \frac{\hat{V}_{RMS}^2}{Z_0} = \frac{V_{MAX}^2}{2 \cdot Z_0} = \frac{\left(E_{MAX} \cdot r_I \cdot \ln\left({}^{r_O}\!/_{r_I}\right)\right)^2}{\frac{120}{\sqrt{\varepsilon_r}} \cdot \ln\left({}^{r_O}\!/_{r_I}\right)} = \frac{\sqrt{\varepsilon_r} \cdot E_{MAX}^2}{120} \cdot r_I^2 \cdot \ln\left({}^{r_O}\!/_{r_I}\right)$$

To find the maximum value, differentiate and equate the derivative to 0.

$$\frac{\partial \hat{P}_{MEAN}}{\partial r_I} = \left[\frac{\sqrt{\varepsilon_r} \cdot E_{MAX}^2}{120}\right] \cdot \left(r_I^2 \cdot \frac{r_I}{r_O} \cdot \left(\frac{-r_O}{r_I^2}\right) + \ln\left(\frac{r_O}{r_I}\right) \cdot 2r_I\right) = \left[\frac{\sqrt{\varepsilon_r} \cdot E_{MAX}^2}{120}\right] \cdot 2r_I \cdot \left(\ln\left(\frac{r_O}{r_I}\right) - \frac{1}{2}\right)$$



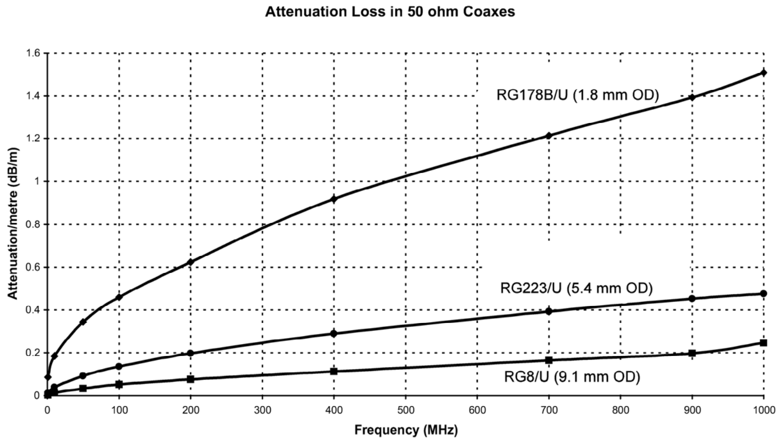
which is zero when

$$\ln\left(\frac{r_O}{r_I}\right)=\frac{1}{2}$$

giving $Z_0\sqrt{\varepsilon_r}=30\,\Omega$ for maximum power transfer. For a useful dielectric such as PTFE (ε_r =2.1), the characteristic impedance for optimum power transfer is therefore 20.7 Ω.

This is for air-spaced coax. For any other dielectric, the curve shifts left.

What coax impedance gives minimum attenuation?



The simple explanation for loss in a coaxial cable is that due to the resistance of the conductors. Since the resistance increases with the square root of frequency due to the ***skin effect***, the simple model shows an attenuation in decibels which increases with the square root of frequency. For this reason larger cable (of a given construction) always has a lower attenuation loss than smaller cable.

The reason why the characteristic impedance comes into the equation is that the attenuation is related to the ratio of the resistance of the cable to its characteristic impedance. For a fixed inner radius of the outer conductor (in other words for a given size of cable), increasing the inner conductor radius makes the impedance lower, but also makes the characteristic impedance lower. Because these two effects do not occur at the same rate, there is an optimum ratio of inner and outer diameters which can be found by mathematical analysis. Ultimately this will yield a ratio of diameters which will give a characteristic impedance (for a given value of dielectric constant).

First set up the rules of this analysis. The cable is assumed to be operating at a frequency where the skin depth is at least several times smaller than the material thickness. This will be reasonable for frequencies above 100 kHz. The loss in the dielectric can be ignored when using an air-spaced coax, or a construction where the dielectric support of the centre conductor is minimised.

In terms of resistance, the total path consists of the outer sheath and the inner conductor. The approximate resistances of the inner and outer rings are:

$$R_I=\frac{\rho\cdot L}{A}=\frac{\rho\cdot L}{2\pi\cdot r_I\cdot\delta}\qquad R_O=\frac{\rho\cdot L}{A}=\frac{\rho\cdot L}{2\pi\cdot r_O\cdot\delta}$$

δ= skin depth. This gives a combined resistance

$$R=R_I+R_O=\frac{\rho\cdot L}{2\pi\cdot\delta}\cdot\left(\frac{1}{r_I}+\frac{1}{r_O}\right)$$

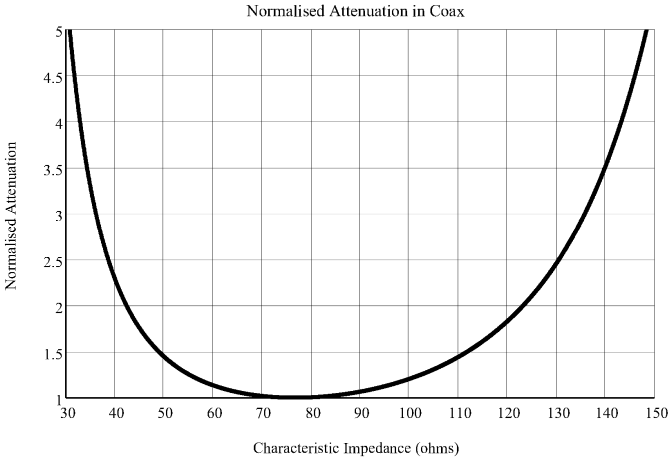
For minimum attenuation, maximise the ratio of Z_0/R ;

$$x = \frac{Z_O}{R} = \frac{60 \cdot 2\pi \cdot \delta \cdot}{\rho \cdot L \cdot \sqrt{\varepsilon_r}} \cdot \frac{\ln\left({}^{r_O}\!/\!{}_{r_I}\right)}{{}^{1}\!/\!{}_{r_I} + {}^{1}\!/\!{}_{r_O}}$$

Take r_O as constant.

$$\frac{\partial x}{\partial r_I} = \frac{60 \cdot 2\pi \cdot \delta}{\rho \cdot L \cdot \sqrt{\varepsilon_r}} \cdot \frac{\left[\left(\frac{r_I}{r_O}\right) \cdot \left(-\frac{r_O}{r_I^2}\right) \cdot \left(\frac{1}{r_I} + \frac{1}{r_O}\right) - \ln\left(\frac{r_O}{r_I}\right) \cdot \left(-\frac{1}{r_I^2}\right)\right]}{\left(\frac{1}{r_I} + \frac{1}{r_O}\right)^2}$$

This is a maximum when $\ln\left(\frac{r_O}{r_I}\right) = 1 + \frac{r_I}{r_O}$, which occurs when the radius ratio is 3.59, giving an impedance of 76.7 Ω (for an air dielectric, ε_r = 1). With a solid dielectric of relative permittivity 2.3, the optimum impedance is ≈50 Ω.



Note that in practice the coaxial cable diameter cannot be increased without limit for high frequency signals. If the frequency gets too high for a given coaxial cable, the transfer mechanism within the cable changes from being only TEM to containing the TE mode. (see *waveguide mode* in the glossary).This higher mode travels at a different speed and therefore wrecks the signal handling qualities of the cable.

How long do you need to measure to get an accurate mean reading?

The mean value of some time varying voltage *v(t)* is simply the 'DC value'. Mathematically this would be written as:

$$\langle v(t)\rangle \equiv \overline{v(t)} = \frac{1}{T} \cdot \int_0^T v(t) \cdot dt$$

Note that there are two different notations for a (time) average {mean}; the angled brackets and the bar. Both are commonly used.

If the function *v(t)* is periodic, and *T* is not an integer number of cycles, there will be an error. This error can be made arbitrarily small by measuring over a longer interval.

If the function has a cycle-mean of 0, "<0.1% error" is not a workable measure. A more practical error measure would be "<0.1% of the peak value". How long should you measure to guarantee such an uncertainty?

When T is larger than 1/*f* , it is not obvious which starting and ending positions give the worst error; mathematical analysis is necessary.

$$m(t) = \frac{1}{T} \cdot \int_0^T V \cdot \sin(2\pi f t + \phi) \cdot dt = \frac{V}{T} \cdot \left[\frac{1}{2\pi f} \cdot \cos(2\pi f t + \phi)\right]_0^T$$

$$\therefore m(t) = \frac{V}{2\pi f T} \cdot \left[\cos(2\pi f \cdot T + \phi) - \cos(\phi)\right]$$

It is convenient to express the value of T in terms of an integer number of cycles plus an offset δ, in radians. The term $2\pi f T$ within the cosine expression therefore simplifies giving:

$$m(t) = \frac{V}{2\pi f T} \cdot [\cos(\delta + \phi) - \cos(\phi)]$$

The largest value in the square brackets is +2, achieved by making $\cos(\phi) = -1 \quad (\phi = \pi)$ and $\cos(\delta + \phi) = 1 \quad (\delta = \pi)$. The worst offset on the mean value is therefore: $|offset| \le \frac{V}{\pi f \cdot T} \le \frac{V}{n\pi}$ where *n* is the number of complete cycles of the mean. However, this formula is overly pessimistic because the worst case occurs roughly in the middle of a cycle.

Worst offset on mean, $\boxed{|offset| \le \frac{V}{(\mathrm{n} + 0.4)\pi}}$

Since any periodic waveform can be reduced to a Fourier series of harmonics, and since the error in the harmonics reduces faster than the error in the fundamental, this equation is a worst case limit for any periodic waveform.

To correctly measure the mean value of a periodic waveform to an uncertainty of better than 0.1% of the peak value, it is best to measure over an integer number of cycles. Failing that, take the mean starting from the peak of the waveform and average over at least 160 cycles, the offset formula being $\boxed{|offset| \le \frac{V}{2n\pi}}$. For an unrestricted starting point for the mean, average over at least 318 cycles.

Measuring over a limited time interval, T_m, reduces the low frequency signal and noise within that interval. The interval creates a high-pass filter with a 3 dB corner at $\frac{1}{4 \times T_m}$, although the initial slope is faster than a single-pole response. Sampling over a 1 second interval, for example, rolls off noise below 0.25 Hz.

How long do you need to measure to get an accurate RMS reading?

For simplicity a sine wave will be considered. The RMS value will be most accurate when taken over an integer number of cycles. (Hence “cycle-RMS” measurements on some scopes.)

$$RMS = \sqrt{\frac{1}{T}\int_0^T V^2 \cdot \sin^2(2\pi f t + \phi) \cdot dt} = \sqrt{\frac{V^2}{T}\left[\frac{t}{2} + \frac{\phi}{4\pi f} - \frac{1}{8\pi f}\sin(4\pi f t + 2\phi)\right]_0^T}$$

$$\therefore RMS = \sqrt{\frac{V^2}{T}\left(\frac{T}{2} + \frac{1}{8\pi f}[\sin(2\phi) - \sin(4\pi f T + 2\phi)]\right)} \le \frac{V}{\sqrt{2}}\sqrt{1 + \frac{1}{2\pi n}} \le \frac{V}{\sqrt{2}}\left(1 + \frac{1}{4\pi n}\right)$$

The sine terms have been maximised and *n* cycles plus a non-integer part have been measured. The true RMS value is $\frac{V}{\sqrt{2}}$, giving $\boxed{\text{uncertainty} \le \frac{1}{4\pi n}}$ per-unit.

To guarantee less than 0.1% uncertainty in the RMS reading when a non-integer number of cycles is used, at least 80 cycles must be measured. However, a factor of two reduction in the uncertainty is achieved by starting the measurement at a zero crossing point.

For rectangular waveforms see EX 14.6.7

KEY ANSWERS

Full answers to the other questions are given at **www.logbook.freeserve.co.uk/seekrets**

ANS 3.4.1:

A) The answer is ±7% by just adding the tolerances. If you were a bit 'over-enthusiastic' then maybe you did this:

$1.01\times1.02\times1.01\times1.03=1.072$ and $0.99\times0.98\times0.99\times0.97=0.932$

This means +7.2% to –6.6%. In many respects the ±7% answer is a better answer; it was achieved with less effort and therefore cost less. The accuracy improvement in doing the calculation "correctly" is usually not justified.

B) $\sqrt{1^2+2^2+1^2+3^2}=\pm3.9\%$: That is what the maths says. Now see if it makes any sense! Components often come in batches where they are all off from nominal in one particular direction. If the 3% components are all high and towards their limit then there is a very high probability (certainly worse than a 25% probability for a flat tolerance distribution) that the 2% component will push the result outside of the 3.9% limit.

ANS 4.5.1:

No single numeric answer is adequate. The 'statistical independence' was irrelevant since a worst case was asked for. You haven't been given the resistor ratio for the divider. Without it you can't work out the actual error, but you can work out a formula for the error. Write down the transfer equation using R1 as the upper resistor and R2 as the lower resistor:

$$\boxed{V_o=V_{IN}\cdot\frac{\mathrm{R2}}{\mathrm{R1}+\mathrm{R2}}=\frac{V_{IN}}{\left(1+\frac{\mathrm{R1}}{\mathrm{R2}}\right)}}$$

You can see by inspection that if R2 >> R1 then there is hardly any error, whereas if R1 >> R2, you 'see' most of the ratio error between the two resistors. The ratio error could be due to TC or long term drift, it doesn't matter. The error amount is multiplied by the R1/R2 ratio, and then divided by (1 + R1/R2) to get its proper weighting factor. This simplifies to :

$$\boxed{TotalError=\frac{RatioError}{\left(1+\frac{\mathrm{R2}}{\mathrm{R1}}\right)}}$$

It is as if the error signal were being injected from the bottom of the network, with the input grounded in terms of this error attenuation. This comes down to a simple rule. If the gain of the network is g (g<1) then $\boxed{TotalError=(1-g)\times RatioError}$

With ±1% resistors the error is hardly anything when the attenuation is hardly anything. When the attenuation is greater than 5, the error is ±2%. At the mid-position the error is ±1%. The error can never by greater than ±2% for this attenuator, regardless of the attenuation factor.

ANS 6.2.1:

Think of the two capacitors in series. The interface between the two insulators is an equipotential and could therefore be replaced by a thin conducting plate.

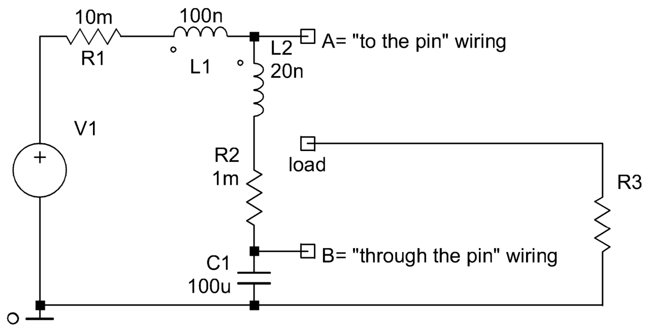
For the capacitor with the dielectric, $C_1=\dfrac{\varepsilon_o\cdot K\cdot A}{G\times d}$.

The other capacitor has an air dielectric, $C_2=\dfrac{\varepsilon_o\cdot A}{(1-G)\times d}$.

The total capacitance is

$$\boxed{C_T=\frac{1}{\left(\frac{1}{C_1}+\frac{1}{C_2}\right)}=\frac{1}{\frac{G\times d}{\varepsilon_0\cdot K\cdot A}+\frac{(1-G)\times d}{\varepsilon_0\cdot A}}=\frac{\varepsilon_0\cdot A}{d}\left(\frac{1}{1-G+\frac{G}{K}}\right)}$$

ANS 6.7.2:



As usual, a correct equivalent circuit makes the solution obvious. If you wire to point A the track length to the capacitor gives significant extra series impedance; the capacitor is less effective at decoupling noise originating at the power supply, V1, or caused by noise in the load. Take the power to the load directly from the pin of the capacitor as illustrated (point B). This gives a minimum of extra series impedance.

Note that this equivalent circuit is shown only for the positive power rail. It should be obvious that this routing is equally important on the ground [0 V] track as well. You could easily make the power supply noise 10× greater than it needs to be by routing the tracks incorrectly to your decoupling capacitors. There is no cost associated with this technique, except perhaps a more difficult PCB layout.

ANS 7.9.1:

$$H=\frac{I}{2\pi r}-\frac{I}{2\pi(r+d)}=\frac{I}{2\pi}\left(\frac{1}{r}-\frac{1}{r+d}\right)=\frac{I}{2\pi}\left(\frac{r+d-r}{r(r+d)}\right)=\frac{I\cdot d}{2\pi r(r+d)}$$

When $r >> d$ this formula reduces to $H=\frac{I\cdot d}{2\pi r^2}$. The field intensity drops as an inverse square law.

ANS 7.9.2:

The formula is $\mathbf{B}=\mu\mathbf{H}$, where $\mu_0=4\pi\times10^{-7}$ H/m is the permeability of free space.

1000 A/m of **H** therefore corresponds to 1.26 mT of **B**.

ANS 7.9.3:

Faraday's law of induction is used, $V=N\cdot\frac{d\phi}{dt}=N\cdot\frac{d(B\cdot A)}{dt}=N\cdot A\cdot\frac{dB}{dt}$

In this case the number of turns, *N*, is one and the loop area, *A*, is 1 cm^2 =10^{-4} m^2.

Also $B=\hat{B}\cdot\sin(\omega t)$; therefore $\frac{dB}{dt}=\omega\hat{B}\cdot\cos(\omega t)$

In terms of the H-field, the induced voltage in the single turn 1 cm^2 loop is:

$$V=10^{-4}\times4\pi\times10^{-7}\times2\pi f\times H=0.79\times10^{-9}\times f\times H$$

ANS 10.2.1:

Current into R1 equals the current out through R2: $\frac{V_{IN}}{R_1}=\frac{-p\cdot V_{OUT}}{R_2}$ $\boxed{\therefore\frac{V_{OUT}}{V_{IN}}=\frac{-R_2}{p\cdot R_1}}$

ANS 10.2.2:

Working from first principles, remember that for linear operation the voltage on the non-inverting input is the same as that on the inverting input. The input is considered as being short-circuited to ground. Using V_N as the noise source feeding the non-inverting input gives the nodal equation:

$\frac{V_N}{R_1} = \frac{p \cdot V_{OUT} - V_N}{R_2}$. Collecting terms, $V_N\left[\frac{R_2}{R_1}+1\right] = p \cdot V_{OUT}$ $\boxed{\therefore \text{Noise Gain} = \frac{V_{OUT}}{V_N} = \frac{1}{p}\left(\frac{R_2}{R_1}+1\right)}$

ANS 10.2.3:

The noise gain referred to the input is: $\frac{V_{IN}}{V_N} = \frac{V_{IN}}{V_{OUT}} \cdot \frac{V_{OUT}}{V_N} = \frac{pR_1}{R_2} \times \frac{1}{p}\left(\frac{R_2}{R_1}+1\right) = 1 + \frac{R_1}{R_2}$

It is usual to neglect the negative sign in the formula, since for noise it has no meaning. The sign is only relevant for input offset TC direction and that is often not guaranteed either.

ANS 10.7.4:

As before, write down the system response by inspection. Look at the top opamp and relate everything to its inputs. The common-mode gain must be nominally zero. Hence the common-mode signal from the two inputs must cancel. Use linear superposition to make the working easy.

A) $V_{OUT} = V_C\left(1+\frac{R_1}{R_2}\right) - V_C\left(1+\frac{R_3}{R_4}\right) \times \frac{R_1}{R_2}$.

For cancellation of the nominal common-mode gain: $1+\frac{R_1}{R_2} = \left(1+\frac{R_3}{R_4}\right) \times \frac{R_1}{R_2}$

$\therefore \frac{R_2}{R_1} + 1 = 1 + \frac{R_3}{R_4}$, giving the matching criterion as $\frac{R_2}{R_1} = \frac{R_3}{R_4}$.

B) A simple circuit simulation using opamp macro-models shows that this scheme can be 30 dB worse on CMRR than the three-opamp configuration, even at frequencies well below the bandwidth of the amplifiers. In this circuit the amplifiers need not be matched. To get any sort of performance at all, the requirements on opamp bandwidth are much greater than for the three opamp configuration. The problem is that phase shift of the common-mode signal through the bottom amplifier means that the common-mode signal does not null properly. This phase shift can be corrected to a limited extent with a capacitor across R2. This trick can give a >20 dB improvement of CMRR in the kilohertz region according to macro-model simulations.

The three-opamp configuration will always give better CMRR performance for any given opamps. Both circuits can be tweaked by the subtle additions of a capacitor and by mismatching a resistor from nominal to account for opamp imperfections.

ANS 10.7.6:

The ideal differential amplifier will not load the low-pass filter since its input impedance will be infinite.

The transfer function of the low-pass filter is $T = \frac{1}{1+j\frac{f}{B}}$.

The error signal for a 1 V input is therefore $V_{Error} = 1 - \frac{1}{1+j\frac{f}{B}} = \frac{1+j\frac{f}{B}-1}{1+j\frac{f}{B}} = \frac{j\frac{f}{B}}{1+j\frac{f}{B}}$

Only the error signal magnitude is important, $|V_{ERROR}| = \frac{f}{B} \cdot \frac{1}{\sqrt{1+\left(\frac{f}{B}\right)^2}} \approx \frac{f}{B} \cdot \left(1 - \frac{1}{2}\left(\frac{f}{B}\right)^2\right) \approx \frac{f}{B}$

$\boxed{CMRR_{dB} \approx 20 \times \log_{10}\left(\frac{B}{f}\right) \text{ dB}}$ for $f < \frac{B}{3}$ A) 40 dB B) 60 dB.

ANS 10.7.7:

Dividing down the input before application to the differential amplifier is unfortunate, but necessary for signals greater than around ±12V.

$$DM\ signal = \frac{V_D}{2} \times G + \frac{V_D}{2} \times G(1-\delta) \qquad DM\ gain = G\left(1 - \frac{\delta}{2}\right)$$

$$CM\ signal = V_C \times G - V_C \times G(1-\delta) = \delta \cdot GV_C \qquad CM\ gain = \delta \cdot G$$

$$\boxed{CMRR = \frac{DM\ gain}{CM\ gain} = \frac{1 - \frac{\delta}{2}}{\delta} \approx \frac{1}{\delta}}$$

To get a CMRR in excess of 80 dB the two attenuators are required to be matched and maintained to better than 100ppm (0.01%).

ANS 11.5.3:

A) There are two effects, both of which make the distortion worse. Firstly an amplifier has a non-linear output resistance. The lower the load resistor the more the signal is dropped across the non-linear output resistance and therefore the larger the distortion. Secondly, as the load increases, the forward gain of the amplifier is reduced. The reduction of distortion due to feedback is reduced when the loop-gain is reduced.

B) The load resistor reduces the forward gain of the amplifier, and it does so without giving an additional phase shift. Less forward gain, with the same phase, always gives increased stability margins. Think of the Bode plot of the amplifier. Halving the forward gain, without introducing a phase shift, increases the gain margin by 6 dB.

ANS 14.6.6:

A) The mean power in a resistor caused by this rectangular waveform is $\overline{P} = D \times \frac{V^2}{R} = \frac{\left(V\sqrt{D}\right)^2}{R}$. The RMS value, described in terms of its effective heating effect, is $V\sqrt{D}$. Working it out the long way:

$$V_{RMS} = \sqrt{\frac{1}{T}\int_0^T v(t)^2 \cdot dt} = \sqrt{\frac{1}{T}\int_0^{DT} V^2 \cdot dt} = \sqrt{\frac{1}{T}\left[V^2 \cdot t\right]_o^{DT}} = \sqrt{\frac{1}{T}V^2 \cdot DT} = V\sqrt{D}\ .$$

B) A positive value and a negative value with the same magnitude both give the same result when squared. Thus regardless of the duty cycle, the RMS value is just V. Mathematically:

$$V_{RMS}^2 = \frac{1}{T}\left[\int_0^{DT} V^2 \cdot dt + \int_{DT}^{T} (-V)^2 \cdot dt\right] = \frac{1}{T}\int_0^T V^2 \cdot dt = \frac{V^2}{T} \cdot [t]_0^T = V^2\ , \qquad V_{RMS} = V$$

C) $V_{RMS}^2 = \frac{1}{T}\left[\int_0^{DT} V^2 \cdot dt + \int_{DT}^{T} (-sV)^2 \cdot dt\right] = \frac{V^2}{T}\left[DT + (T - DT)s^2\right] = V^2\left(D + (1-D)s^2\right)$

$\boxed{V_{RMS} = V\sqrt{D + (1-D)s^2}}$ s=0 gives answer A. s= ±1 gives answer B.

D) From answer C it should be clear that the answer to D is the same as to C.

E) AC coupling the waveform is equivalent to subtracting the mean value from the waveform.

$$\overline{V} = \frac{1}{T}\int_0^T v(t) \cdot dt = \frac{1}{T}\int_0^{DT} V \cdot dt = \frac{V}{T} \cdot DT = DV$$

In terms of the equation of section C, the high value is $V - DV = V(1-D))$ and the low value is $-DV$, with the duty cycle remaining the same. $s = \frac{DV}{V(1-D)} = \frac{D}{1-D}$.

Then $V_{RMS} = V(1-D)\sqrt{D + (1-D)\frac{D^2}{(1-D)^2}} = V(1-D)\sqrt{D + \frac{D^2}{1-D}} = V(1-D)\sqrt{\frac{D - D^2 + D^2}{1-D}}$

$\boxed{V_{RMS} = V\sqrt{D(1-D)}}$ For an AC coupled square wave (D=0.5), the RMS voltage is ptp/2.

ANS 16.4.3: see the Appendix: "How can you convert resistive loads to VSWR and vice versa?"

ANS 17.11.5:

The input impedance of the network can be written from inspection, using the "product over sum" rule for parallel impedances.

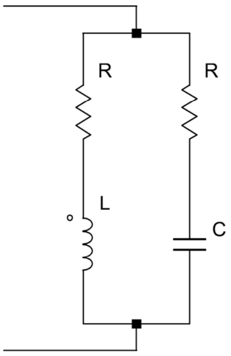
$$Z_{IN} = \frac{R_0 \cdot \frac{1}{j\omega C}}{R_0 + \frac{1}{j\omega C}} + \frac{R_1 \cdot j\omega L}{R_1 + j\omega L} = \frac{R_0}{1 + j\omega CR_0} + \frac{j\omega L}{1 + j\omega \frac{L}{R_1}}$$

In order for this network to have a VSWR of 1 at all frequencies, the input impedance has to be R$_0$ at all frequencies. The possibility of achieving this is not immediately obvious. However by making $CR_0 = \frac{L}{R_1}$ the two terms for input impedance share a common denominator.

$$Z_{IN} = \frac{R_0 + j\omega L}{1 + j\omega CR_0} = R_0 \times \frac{1 + j\omega \frac{L}{R_0}}{1 + j\omega CR_0}$$

It is now clear that Z$_{IN}$ will be the required value if, and only if, $\frac{L}{R_0} = \frac{L}{R_1} = CR_0$.

Thus $\boxed{R_1 = R_0}$ and $\boxed{L = CR_0^2}$.



There is in fact no need for the link between the node "out" and the series RC chain. This link can be removed without affecting the performance. This idea is useful when the intermediate nodes are not accessible, as would be the case when some of the circuit elements are parasitic. Having removed the link between the two series chains, the order of the series elements can be swapped, if desired, and an interesting phase compensation network results. The network presents an input impedance of R at all frequencies. Both the original network in the exercise and this new network are sometimes referred to as Zobel networks.

ANS 18.4.1:

A) Four? You were expecting *conduction*, *convection* and *radiation*. Forced convection and natural convection only count as one. Mechanical engineers and thermodynamics specialists seem to ignore heat transportation by electric current. This is quite an omission because ***Peltier*** coolers (also known as TECs= Thermo-Electric Coolers), whilst quite expensive ($30), are certainly effective devices. This transport mechanism cannot be dismissed as conduction because it is electrically controllable.

B) Trick question. You cannot state what the dominant mode is because it depends on the system concerned. Any of the modes could be dominant in this temperature range.

INDEX