EE414 Transceiver Design Lab

Home

Announcements

Handouts

Useful Information

EE414 - Design of RF Integrated Circuits for Communications Systems

Professor Tom Lee

Department of Electrical Engineering

Spring 2000/2001

MW 12:50 - 2:05 pm
550-553R

Go To Spring 2002 Website

Course Description

This course covers the design, construction, analysis and experimental evaluation of radio-frequency circuits at the transistor level, with a focus on microstrip implementations in low GHz range. Throughout, there is an exceptionally strong emphasis on reconciliation of theory with experiment. Students will design, construct, and experimentally characterize every block in a 1-GHz transceiver, including antennas, low noise amplifiers and modulator/demodulators, roughly at the rate of one significant block per week. Performance will be evaluated using equipment such as noise figure meter, phase noise analyzers, spectrum analyzers, vector network analyzers, and time-domain reflectometers. The course culminates in groups of students successfully demonstrating two-way wireless communications with their hardware. Prerequisites: 314, 344.

Instructor

Professor Tom Lee tomlee@ee

Office hours: generally the hour after class, or by appointment
CIS-205 (725-3709)

Teaching Assistant

Moon Jung Kim ta414@smirc
Office hours: any time at CIS, or by appointment
CIS-063 (725-4538)

Course Administrative Assistant

Ann Guerra guerra@par
CIS-207 (725-3725)

Lab location

Packard 002 (725-1769)

Handouts

- HO #1 EE414: Transceiver Design Laboratory
- HO #2 Instrument Reference Manuals and Component Data Sheets
    - Supplimentary Handout from EE344:
      Quick Reference Guide: Using the Digitizing Scope to Make TDR Measurements

- HO #3 Lab Assignment 1: Microstrip Structures
- HO #4 Connectors, Cables and Waveguides
- HO #5 Microstrip, Stripline and Planar Passive Elements

HO #1 ~ #5 are available as hardcopies in CIS-070.

- HO #6 Passive Components
- HO #7 Time Domain Reflectometry
- HO #8 Network Analyzers
- HO #9 Derivation of Fringing Correction
- HO #10 Noise Figure Measurement
- HO #11 Narrowband LNA Design Lab
- HO #12 Narrowband LNA Design
- HO #13 Antennas
- HO #14 Filters
- HO #15 Microstrip Filters
- HO #16 Antenna and Filter Design Lab
- HO #17 Power Amplifier Design Lab
- HO #18 Synthesizer Design Lab
- HO #19 MC12181 Datasheet

Quick Reference Guide: Using the Digitizing Scope to Make TDR Measurements

The HP 54120B digitizing oscilloscope in the lab can be used as a time-domain reflectometer. This guide takes you through the calibration and introduces some of the basic features of the scope.

As with other equipment in the lab, the HP 54124A (sampling head) is static-sensitive, so be sure to ground yourself to the chassis before connecting or removing your circuits.

Note: The softkeys at the bottom of the display are referred to as "menu keys," and the softkeys along the right side of the display are "function keys."

Calibration:

1. Put the scope in the *persist* mode

   Select `Display` from the bottom menu keys (you may need to press `More` to see all the options).

   Press the `Display Mode` function key until the setting reads `Persist`.

   Ensure the persist setting is 300ms.

2. Perform Calibration Sequence

   Select (`More`) `Network` from the bottom menu.

   Select `Cal` on the top function key.

   Press `Preset Reflect Channel`.

   Press `Reflect Cal`.

   Connect the short from the BNC calibration kit.

   Press `Reflect Cal`.

   Connect the load termination from the BNC cal kit.

   Press `Reflect Cal`.

Make measurements:

1. (Optional) Adjustments to the display

   To stretch the time axis, select `Timebase` from the menu keys. Then use `⇑⇒` to stretch the axis.

   To adjust the vertical position, select `Channels` from the menu keys, `Offset` from the function keys.

2. Use the cursor to read the display

   Go back to the network mode (select `Network` from the menu keys).

This time, select [Reflect] on the top funcion key.

Hit the [Cursor] function key.

Turn the dial to the desired point in the waveform to make measurements of reflection coefficient (ρ), characteristic impedance (Z), time (Δt), and distance (d). *Be sure that the measurements make sense.*

Plot:

1. Generate plot

   Select [Plot] from the menu keys.

   Make sure [Auto Pen] reads `on`.

   Make sure [Pen Speed] reads `Fast`.

   Generate your plot in three steps by pressing the keys labeled [Plot Waveforms], [Plot Graticule], and [Plot Factors].

2. Abort plot

   Press [Pause], [Continue], *then* [Abort].

Other tips:

Don't set anything heavy (or valuable!) on the front ledge of the cart–it is not very sturdy.

Passive Components

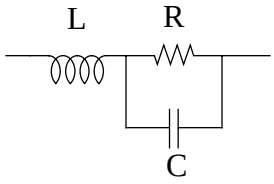
1.0 Introduction

In this chapter, we examine the properties of passive components commonly used in RF work. Because parasitic effects can easily dominate behavior at GHz frequencies, our focus is on the development of simple analytical models for parasitic inductance and capacitance of various discrete components.

2.0 Resistors

Even a component as simple as a resistor exhibits complex behavior at high frequencies. We may construct a very simple model by acknowledging first that current flows in both the connecting leads and the resistor proper. The energy stored in the magnetic field associated with that current implies the presence of some series inductance (typically about 0.5nH/mm for leads in axial packages, as a rough approximation[1]). In addition, there is necessarily some capacitance that shunts the resistor as well, since we have a conductor pair separated by a distance. The simplest (but by no means unique) RF lumped circuit model for a physical resistor might then appear as follows:

FIGURE 1. Simple lumped RF resistor model



The presence of parasitic inductance and capacitance causes the impedance to depart from a pure, frequency-independent resistance. Very low values of resistance suffer from an early impedance increase, starting approximately at a frequency where the reactance of the series inductance becomes significant compared with the resistance. Similarly, high resistances suffer a premature impedance decrease from the shunt capacitance. The frequency range over which the impedance remains roughly constant (at least for our simple model) is maximized for some intermediate (and definite) resistance value. As one might suspect from transmission line theory, this magic value is simply given by

$$R_{opt} = \sqrt{\frac{L}{C}} = Z_0. \tag{1}$$

1. For various equations for inductance and capacitance, see Chapter 2 of T. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge University Press, 1998.

The formal derivation of Eqn. 1 (which we will not carry out here) begins by writing an expression for the impedance magnitude, and then solving for the condition of maximal flatness by maximizing the number of derivatives whose value is zero at zero frequency[2]. If the resistor value is smaller than Z_0, the impedance of our model only rises, with a radian corner frequency given approximately by R/L. If the resistance exceeds Z_0, the impedance initially drops, with a corner frequency of roughly $1/RC$. If the resistance equals Z_0, the bandwidth f_{max} over which the impedance remains approximately constant is given by the resonant frequency of the LC combination. The smaller the LC product, the greater the frequency range over which the resistor looks approximately resistive.

For reference, typical model parameter values for some representative resistors are given in the following table. Note that the maximally-flat impedance levels are in the range of 100-200Ω. That is yet another reason why transmission line impedances tend not to be far removed from that range of values.

TABLE 1. Approximate element values for simple lumped RF resistor model

Resistor type	L	C	Z_0	f_{max}
0.5W axial-lead (3cm total length)	15nH	0.5pF	170Ω	1.8GHz
0.25W axial-lead (2cm total length)	10nH	0.25pF	200Ω	3.2GHz
Type 0805 surface mount	1nH	0.06pF	130Ω	20GHz

Note that the values for Z_0 are all in excess of typical line impedances (such as 50Ω). If, for example, we are to provide terminations for a 50Ω line, then the largest bandwidth is obtained with a parallel combination of several devices, rather than with a single 50Ω resistor. Four 1/4W resistors of 200 ohms each will do a very good job of providing a 50Ω termination over a bandwidth in excess of 1GHz. Similarly, a parallel combination of 0805 surface mount resistors will provide an excellent termination over a bandwidth in excess of 10GHz. The table also shows why conventional resistors (particularly the 1/2W variety) are rarely used in microwave work. If higher power terminations are required, it is preferable to make them out of parallel combinations of lower-power, higher-frequency resistors for operation over the largest bandwidth.

As an aside, the numerical identifiers for surface mount components convey something about the physical dimensions. Within truncation errors, the first pair of numbers is four

2. Lee, op. cit. As mentioned there, it is often easier to carry out this procedure on the square of the magnitude.

times the length in millimeters, while the second pair is four times the width. Some common sizes are shown in the following table, along with their power dissipation levels:

TABLE 2. Some surface mount resistor packages and characteristics

Package	Dimensions (mm x mm)	P_{diss} (mW)	Approx. min. capacitance (pF)
0402	1x0.5	60	0.03
0603	1.6x0.8	60	0.05
0805	2x1.25	100	0.06
1206	3.2x1.6	125	0.09
1210	3.2x2.5	250	0.12
1812	4.5x3.2	500	0.16
2512	6.4x3.2	1W	0.18

To convey an idea of the parasitics associated with these packages, consider the largest size listed in the table, the 2512, whose 6.4x3.2x3.2 dimensions are associated with ~2.5nH parasitic inductance and ~0.18pF shunt capacitance. If measured values are unavailable, one may estimate the series inductance for the other packages with the aid of the following equation for the inductance of an infinitesimally thin flat sheet:

$$L_{sheet} \approx \frac{\mu l}{2\pi}\left[0.5 ln\left(\frac{2l}{w}\right) + \frac{w}{3l}\right]. \quad (2)$$

This formula is appropriate because the resistive material is almost always a flat, thin layer deposited on the top surface of a much thicker insulating substrate (even for “thick film” resistors). For the aspect ratios typically encountered in low-power surface mount components, the inductance is usually in the range of 0.2-0.5nH/mm.

One uses this formula twice to estimate the total parasitic inductance. One component of the total inductance is that of the main body, so its length and width are first plugged into the equation. To this (usually) dominant term, one must also add the inductance due to the flat vertical portions that contact the ends of component, with the height now replacing the width in Eqn. 2. The solder meniscus effectively thickens these vertical stubs, however, so it isn’t quite fair to use the full inductance of each vertical section. As an arbitrary choice, 1/2 to 2/3 of the computed value of the vertical stubs is not an unreasonable factor. Using the former factor, we estimate an inductance of about 2.6-2.7nH for the 2512 package, in reasonably good agreement with measurements.

Note that the inductance per length here is considerably lower than the 1nH/mm rule of thumb that typically applies to round component leads. The reason is that the thick and

wide shape of the surface mount components spreads out the magnetic field lines, thereby reducing flux density and, hence, inductance.

Estimating the capacitance of this type of structure is somewhat complicated, both because of the dominance of fringing and because the value when mounted on a PC board will generally differ from that measured in isolation with no other dielectrics nearby. Compounding this difficulty is the variability of the dielectric material which forms the body of the resistor. Despite all of these issues, we can offer a cheesy approximation based on the formula for the capacitance per length of a dipole antenna made out of a cylindrical conductor (see the chapter on antennas for a derivation):

$$C \approx \frac{1}{c^2 \frac{\mu_0}{2\pi}\left[ln\left(\frac{2l}{r}\right) - 0.75 \right]} \approx \frac{2\pi\varepsilon_0}{ln\left(\frac{2l}{r}\right) - 0.75} \approx \frac{5.56\times10^{-11}}{ln\left(\frac{2l}{r}\right) - 0.75}. \quad (3)$$

This equation yields an estimate of the capacitance (per length) between ground and a conductor of length *l* and radius *r*. Table 2 provides approximate minimum shunt capacitance values based on this simple (and admittedly simple-minded) equation. These values are a rough lower bound, and assume (among other things) a unit dielectric constant and no additional metal pads, etc. The capacitances will be boosted by the dielectric constant of the package, as well as that of FR4. Even so, the values in Table 2 are not terribly far off from values typically observed experimentally (typically within a factor of 1.5), because external fringing dominates.

This approximate method may be used to estimate package parasitic capacitances of surface mount inductors, as well as those of ordinary components of circular cross-section. Just remember that package parasitics may account for only a part of the total; some parasitics may arise internally (e.g., turn to turn winding capacitance in inductors). The computed package capacitance is therefore an estimate of a lower bound on the parasitic capacitance.

As a final comment, note that many axial-lead resistors are based on a carbon composition, which consists of a resistive powder formed into a cylindrical shape. Unfortunately such resistors can exhibit significant 1/*f* noise, with a power spectral density proportional to the dc bias current flowing through the resistor. Carbon film resistors are substantially better in this regard, while metal film resistors are even better. Although the 1/*f* corners are generally well below the RF range, one must be aware that oscillators can upconvert low frequency noise into phase noise near the carrier.[3] Thus, even though 1/*f* noise is usually not an issue in circuits such as RF amplifiers, it cannot be completely neglected in all RF circuits. Fortunately, surface-mount resistors are generally of the film variety.

3. Lee, op. cit.

3.0 Capacitors

Many different dielectric materials are used in an effort to satisfy the numerous conflicting demands made on capacitor performance. Trade-offs among breakdown voltage, temperature coefficient, RF loss, and capacitance density inevitably lead to the many types of capacitors presently available. Space does not permit an encyclopedic review of all capacitor types, so we focus only on those that are commonly encountered in high frequency circuits.

The lowest-loss capacitors are made with air (or vacuum) as the dielectric. Higher densities with low loss may be obtained with mica (a naturally occurring mineral) and polystyrene. Although polystyrene has excellent electrical properties, it possesses an unfortunately low melting point, which limits use to temperatures below 85°C. One must consequently exercise great care in soldering. PTFE is also an exceptionally low loss dielectric, as noted earlier. Because of the expense of fabricating good thin films of PTFE, however, capacitors made with it tend to have rather large dielectric thicknesses, leading to low capacitance densities (but very high breakdown voltages).

Ceramic capacitors themselves come in a number of varieties, distinguished by the characteristics of their dielectrics. To keep track of the many permutations, the Electronics Industry Association has settled on a three-character nomenclature. The first character is a letter that indicates the minimum operating temperature. The second character is a number that indicates the maximum temperature, and the third character is a letter that conveys the maximum capacitance change over the entire operating temperature range. The particulars are shown in the following table:

TABLE 3. Capacitor codes (EIA)

Min. temp. (°C)	Max. temp. (°C)	Max. % cap. change over temp. range	Max. % cap. change over temp. range
X: –55	3: +45	A: ±1	P: ±10
Y: –30	4: +65	B: ±1.5	R: ±15
Z: +10	5: +85	C: ±2.2	S: ±22
	6: +105	D: ±3.3	T: –33, +22
	7: +125	E: ±4.7	U: –56, +22
		F: ±7.5	V: –82, +22

For example, a capacitor with a designation of X7R exhibits at most a ±15% capacitance variation over an operating temperature range of –55°C to +125°C.

Although a zero temperature coefficient is most commonly desired, there are important instances in which one wants instead a nonzero TC of a specified value. Oscillators are

one example; inductors typically exhibit a positive TC[4], so capacitors possessing a compensating negative TC are needed to produce an oscillation frequency with an overall zero TC. The characteristics of capacitors with controlled TC are identified by the letter N (for "negative") or, more rarely, the letter P (yes, for positive), followed by the maximum TC magnitude in ppm per degree C. A designation of N750 thus represents a capacitor with a –750ppm/°C temperature coefficient. Just to make things confusing, however, there is an alternate system of codes that conveys the same information. Designed to save space for printing on small components, the three-digit EIA code unfortunately does not directly convey numerical information about the actual TC. The following table provides the necessary translation between the two labeling conventions:

TABLE 4. Capacitor TC codes

Older designation	Three-digit EIA	Older designation	Three-digit EIA
NP0	C0G	N330	S2H
N033	S1G	N470	T2H
N075	U1G	N750	U2J
N150	P2G	N1500	P3K
N220	R2G	N2200	R3L

The first letter in the 3-digit TC convention conveys information about the TC's significant digits. The values are a subset of the values of standard resistors. For example, one can discern from the table that P = 1.5, R = 2.2, S = 3.3, T = 4.7, and U = 7.5. The middle digit of the code is the exponent. The NP0 designation (C0G) stands for "negative-positive-zero" and refers to the characteristics of a composite of negative- and positive-TC materials to yield a nominally zero TC (typically, a maximum of ±30ppm/°C).[5] The capacitance thus stays within approximately 0.15% of the nominal value over the military temperature range (–55°C to 125°C). Capacitance values of up to about 10nF are available in the standard surface mount package sizes. The loss of NP0/C0G is the lowest of the standard types, with peak *Q* values in excess of 500-600 at low frequencies. This material also exhibits a low voltage coefficient.

Other commonly used materials include the somewhat less stable, but higher dielectric constant, X7R ceramic. Surface mount types with values up to about 100nF are available. As mentioned earlier the capacitance might vary as much as ±15% over the military temperature range. Unlike C0G, the capacitance decreases (roughly linearly) with increasing DC bias, with up to an additional 30% drop at the rated voltage. This variation with voltage is associated with the piezoelectric nature of the dielectric, and the nonlinear behavior

4. Consider that inductance is dimensionally proportional to length, and that most materials expand when heated. Thus, most physical inductors possess positive TCs.

5. Note that these designations contain the numeral 0 and not the letter O.

can generate significant distortion when these capacitors are used in the signal path. In addition, most X7R formulations are two orders of magnitude lossier than C0G materials.

High-K (high dielectric constant) ceramics, such as Y5V, give us capacitors that are physically the smallest, but which suffer from extremely high TCs (e.g., up to an astounding 80% drop in capacitance at zero bias over a temperature range of –30°C to 85°C), and losses that are a third of X7R. The voltage coefficient is also strongly negative, and one may expect a capacitance drop of up to 75% at the rated voltage. Such capacitances actually make effective mixers, so beware (or exploit this behavior). Furthermore, such dielectrics are piezoelectric to a surprising degree. It is not unusual for a sharp mechanical shock to generate spikes of volts (sometimes many tens of volts). Even if the spike does not cause direct damage to delicate circuitry, it should be obvious that the microphonic behavior of high-K capacitors can lead to a host of objectionable problems, especially if connected to sensitive circuit nodes and subjected to vibration (as in mobile applications). The most common use of these capacitors is therefore as supply bypasses, rather than in the signal path. Values up to about 1μF are available in the standard surface mount packages.

One should not overlook the option of making capacitors with the PC board as the dielectric. It is frequently convenient for trimming purposes to realize some part of a desired capacitance in PC board form to permit adjustment after fabrication. In any case, it's a good idea to be aware of how much capacitance is associated with a given area of conductor, if for no other reason than to estimate layout parasitics. With FR4, one can expect about $5pF/cm^2$ with a 1/32" (0.8mm) thick substrate, or roughly $2.5F/cm^2$ on a 1/16" (1.6mm) substrate. The loss of FR4 is quite tolerable, being modestly better than that of X7R or Y5V. Still lower loss (and somewhat lower capacitance), of course, is obtained with a higher-quality board material, such as PTFE or RO4003. More discussion on the use of PC board traces for realizing capacitances and inductances is found in the chapter on microstrip.

Capacitor values are encoded as three digits stamped somewhere on the body (if the digits fit), followed by a letter that identifies the tolerance (see Table 5). The first two digits are a mantissa, and the third is an exponent. The implicitly understood unit is the picofarad.

TABLE 5. Capacitor tolerance codes (EIA)

Identifier	Tolerance (pF)	Tolerance (%)
B	±0.1	
C	±0.25	
D	±0.5	
E		±25
F	±1	±1
H		±2
J		±5

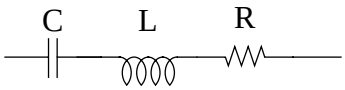
TABLE 5. Capacitor tolerance codes (EIA)

Identifier	Tolerance (pF)	Tolerance (%)
K		±10
M		±20

Hence, “221K” stands for a 220pF capacitor with ±10% tolerance, and “105M” stands for a 1μF, ±20% capacitor. Occasionally, some other conventions are used, but this scheme is by far the most widespread. If in doubt, one can always verify a conjecture with an actual measurement.

Just as with resistors, parasitic effects cannot be ignored at radio frequencies. The simplest lumped RF model for real capacitors includes lead or terminal inductance (as before, this may be estimated as roughly 1nH/mm for round wire leads), and a resistive term to account for losses:

FIGURE 2. Simple lumped capacitor model



The inductance for surface-mount packages can be estimated using Eqn. 2, as before.

The resistive term of the model accounts for the effect of at least two distinct mechanisms. One is the loss of the dielectric, and the other is conductor loss (which is exacerbated at high frequencies by skin effect). Loss is often characterized by either a dissipation factor, *D* (or, equivalently, a loss tangent, tanδ). Dissipation factor is simply the reciprocal of *Q*, while loss tangent is defined as the ratio of the imaginary and real parts of the dielectric constant. Strictly speaking, loss tangent applies only to the dielectric material, but it is often used to include all losses in a capacitor. In this latter case, loss tangent is the same as the capacitor dissipation factor. The reason for these multiple ways of describing loss is cultural. Power electronics folks tend to think in terms of power factor (the cosine of the phase angle between voltage and current, which angle is the same as that of the impedance), RF engineers generally think in terms of *Q*, and materials scientists tend to focus on loss tangent (dissipation factor, *D*).

The definition of power factor means that it is equal to the cosine of the arctangent of *Q* (the proof is left to you, because it is obvious that you don’t have enough fun). For sufficiently large *Q*, the power and dissipation factors converge. For example, a *Q* value in excess of 7 assures an error of less than about 1%. For all capacitors worth using in the signal path, *Q* will certainly be large enough that one may take loss tangent and power factor to be equal in practice.

Given these definitions, the component of effective series resistance (ESR) due to dielectric loss is

$$R \approx \frac{D}{\omega C}. \tag{4}$$

This formula is valid only at frequencies well below the series resonance. Clearly, ESR is a frequency dependent quantity, especially when skin-effect conductor loss is considered as well.

At frequencies well above resonance, the resistance becomes proportional to frequency because the inductive reactance dominates, leading to the following approximation:

$$R \approx D\omega L. \tag{5}$$

We can deduce several important facts from the series *RLC* model. Above the resonant frequency of the network, the combination appears inductive, and the impedance therefore increases with frequency. The minimum impedance is reached at the resonant frequency. If a capacitor is used, say, as a power supply bypass, it is important to recognize that the quality of the bypassing will diminish at higher frequencies because of series inductance. Simply exhibiting inductive behavior need not preclude use, however, since the most relevant quantity is the magnitude of the impedance. If this is sufficiently low, the capacitor can still act satisfactorily as a bypass element, even when operating above the resonant frequency.

As a rough calibration on the magnitudes of these parasitic elements, consider the following table of parameters (at 100MHz):

TABLE 6. Representative capacitors and lumped model parameters at 100MHz

Type	***C***	***L***	***R***	***SRF***
Ceramic disc (C0G/NP0)	10nF	10nH	0.5Ω	17MHz
0805 C0G/NP0	10nF	~1nH	0.03Ω	50MHz
0805 C0G/NP0	100pF	~1nH	0.25Ω	500MHz

In the table, the disc capacitor is assumed to have a total length (measured from the tip of one lead, through the disc body, to the tip of the other lead) of about 10mm. The 100MHz test frequency considerably exceeds the 17MHz self-resonant frequency in this case, so the effective series resistance is due more to the lead inductance, rather than to the intrinsic capacitance. By exerting a little effort to shorten lead length, it is possible both to increase the self-resonant frequency and reduce *R* by modest amounts.

It should be reiterated that loss is a strong function of both frequency and dielectric composition. Thus, the resistance values in Table 6 cannot be treated as universal constants. Your mileage may vary.

4.0 Inductors

We have already incorporated inductance in many of the foregoing equations in a piecemeal manner. We now present a number of additional formulas for commonly-encountered geometries. In all that follows, the equations strictly apply only at DC, unless stated otherwise. At high frequencies, inductance drops because the shrinking of skin depth causes the contribution of internal flux to diminish. Fortunately, internal flux generally accounts for only a small percentage (e.g., below 5%) of the total, so its reduction does not cause dramatic changes in the overall inductance. Nonetheless, it is worthwhile avoiding unpleasant surprises by knowing explicitly what assumptions have gone into the derivations of formulas.

4.1 Flat sheets

We've already presented a formula for the inductance of a current sheet. We repeat it here so that all the inductance formulas are in one place for easy reference:

$$L_{sheet} \approx \frac{\mu l}{2\pi}\left[0.5 ln\left(\frac{2l}{w}\right) + \frac{w}{3l}\right] = 2\times 10^{-7} l\left[0.5 ln\left(\frac{2l}{w}\right) + \frac{w}{3l}\right]. \quad (6)$$

4.2 Wire inductors

It is frequently desirable to know the inductance of lengths of conductor, either because parasitics need to be quantified, or because one desires to use the inductance as a circuit element. If we may neglect the influence of nearby conductors (i.e., if we assume that the return currents are infinitely far away), the DC inductance of a round wire is given by:[6]

$$L \approx \frac{\mu_o l}{2\pi}\left[ln\left(\frac{2l}{r}\right) - 0.75\right] = 2\times 10^{-7} l\left[ln\left(\frac{2l}{r}\right) - 0.75\right]. \quad (7)$$

For a 2mm long standard IC bondwire, this formula yields 2.00nH, leading to an oft-cited rule of thumb that the inductance of thin, round conductors is approximately 1nH/mm. Notice that the inductance does grow faster than linearly with length because there is mutual coupling between parts of the wire (i.e., there is a weak transformer action) with a polarity that aids the inductance. From the logarithmic term, however, we see that this effect is minor. For example, going from 5mm to 10mm changes the DC inductance per mm from 1.19nH to 1.33nH (at least according to Eqn. 7). The inductance is similarly insensitive to the wire diameter, so even the larger conductors found in discrete circuits possess inductances of the same general order (e.g., 0.5 nH/mm).

If there is a conducting plane nearby, defined loosely as closer than a distance approximately equal to the length of the wire, then the inductance will be noticeably lower than

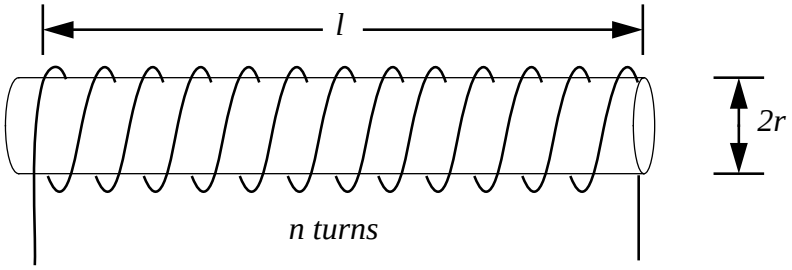
6. *The ARRL Handbook*, American Radio Relay League, 1992, p. 2-18. The proximity of conducting planes may be ignored as long as they are located a distance away equal to one or two lengths, at minimum.

that given by Eqn. 7. Intuitively, this reduction comes about as follows. Current flowing in the wire (which may be thought of, say, as positive charges moving in the x-direction) induces an image current in the ground plane (e.g., negative charges also moving in the x-direction). Opposite charges moving in the same direction are equivalent to two currents flowing in *opposite* directions, so their magnetic fields tend to cancel somewhat, leading to a reduction in magnetic flux. The closer the plane, the more dramatic the reduction in flux (and, therefore, in inductance).

4.3 Air-core solenoids

Although our focus is on components that may be realized in a largely planar universe, more inductance per volume can be obtained with a classic 3-D textbook structure: the single-layer solenoid:

FIGURE 3. Single-layer solenoid



Assuming that, unlike the inductor shown in the figure, the turns are tightly packed ("close-wound"), the inductance in *microhenries* is given by a famous formula presented by Wheeler in the late 1920s:

$$L \approx \frac{n^2 r^2}{9r + 10l}, \tag{8}$$

where r and l are in *inches*.[7] In SI units, the formula is:

$$L \approx \frac{\mu_o n^2 \pi r^2}{l + 0.9r}, \tag{9}$$

where a free-space permeability is assumed. These formulas provide remarkable accuracy (typically better than 1%) for close-wound single-layer coils as long as the length is greater than the radius.[8]

7. H. A. Wheeler, "Simple Inductance Formulas for Radio Coils," *Proceedings of the IRE*, v. 16, no. 10, October, 1928, pp 1398-1400.

8. As discussed later, the best Q is generally obtained when the winding pitch is approximately twice the wire diameter.

For those interested in the origin of this famous and very widely used equation, its derivation begins with the standard undergraduate physics equation for an infinitely long solenoid. For any segment of length *l* of this infinite structure, the inductance in SI units is given by

$$L = \mu_o n^2 \frac{A}{l} = \frac{\mu_o n^2 \pi r^2}{l}, \tag{10}$$

where *A* is the cross-sectional area of the solenoid and *n* is the number of turns contained in the segment under consideration. The important thing to note is that the inductance drops as the solenoid lengthens, all other parameters held constant.

The magnetic field strength along a finite-length solenoid naturally diminishes near the ends. The inductance therefore drops; the solenoid acts as if it were longer than its physical length. If the solenoid is very much longer than its radius, the finiteness is not felt as acutely. Thus, the correction for end effects is a function of the length-to-radius ratio. A famous paper by Nagaoka provides a table and curves for the correction factor, and later work by Grover provides an infinite series which may be truncated as necessary for a given level of accuracy. From these works one may discern that, as a first approximation, a simple estimate for the effective electrical length is the physical length, augmented by the radius. This ad hoc correction is similar to that for the fringing term in capacitors, and turns out to be surprisingly good:

$$L = \frac{\mu_o n^2 \pi r^2}{l + r}. \tag{11}$$

This formula is perfectly respectable, but more rigorous analysis reveals that it does underestimate the inductance slightly[9]. Wheeler's formula does a better job simply by adding 90%, rather than 100%, of the radius to the length.

The effective shunt capacitance across the inductor terminals depends on the boundary conditions to a significant degree. For example, if one terminal is grounded, the effective capacitance is largely independent of the capacitance between adjacent turns. Rather, it depends more on external fringing. This latter capacitance is somewhat difficult to compute analytically. To the best of the author's knowledge, no analytical solution has ever been published. Consequently, the best we can offer here is a semi-empirical formula which assumes that the wire insulation has a relative dielectric constant close to unity, in addition to one grounded terminal. Within the validity of these assumptions, the effective shunt capacitance is approximately[10]

9. Actually, for rather loosely wound coils, the ad hoc approximation actually tends to do a bit better than Wheeler's formula, because flux density dips in the space between the windings. The consequent reduction in inductance is small, but may be accounted for by treating it as an additional effective increase in length.

10. This equation is based loosely on data and a formula due to Medhurst, *Wireless Eng.*, Feb., 1947, pp. 35-43, and March, 1947, pp. 80-92. The coefficients have been chosen to improve accuracy and reduce complexity over Medhurst's formula, as well as to employ SI units.

$$C \approx \pi\varepsilon_0 \left[0.4\left(\frac{l}{D}+1\right) + 0.9\sqrt{D/l} \right] D. \tag{12}$$

This equation matches Medhurst's data within 5% for l/D values ranging from 0.1 to 50.[11] Note that the primary dependence is on the coil diameter, with a weaker dependence on total length. Hence, for a given value of inductance, the highest self-resonant frequencies tend to be obtained with coils possessing the smallest radii. Regrettably, the resulting coil form factor is generally at odds with the goal of maximizing Q.

To obtain accurate estimates of the total shunt capacitance, one must be careful to account also for the capacitance associated with any length of ungrounded lead. For this purpose, one may use the formula for the capacitance of an isolated wire, repeated here for convenience:

$$C \approx \frac{2\pi\varepsilon_0}{ln\left(\frac{2l}{r}\right) - 0.75} \approx \frac{5.56\times10^{-11}}{ln\left(\frac{2l}{r}\right) - 0.75}. \tag{13}$$

One problem with solenoid structures is that they are not self-shielding. Unwanted, and very troublesome, coupling can therefore occur between the inductor and other parts of a circuit, with attendant negative performance implications. Cylindrical shields are thus often placed over such inductors. However, such a shield is uncomfortably similar to (actually, the same as) a shorted single-turn secondary transformer winding. To avoid serious reduction of inductance and Q from induced image currents (also known as eddy currents), the shield's diameter should be at least twice that of the coil (and preferably more) to place the image currents a reasonably large distance away and render their effects negligible.

An alternative is to use a toroidal inductor. Such a structure is magnetically (but not electrostatically) self-shielding if the core material is of sufficiently high permeability. The magnetic flux will then be concentrated in the core, leaving little to leak out. Sadly, all known magnetic core materials are rather lossy at high frequencies, so toroids are widely used only at lower frequencies (e.g., typically well below a few hundred MHz).

Most manufacturers of toroids specify the core's "A_L" value, which they often cite as some number of mH per 1000 turns. Unfortunately, that convention implies a linear dependence of inductance on the number of turns, and this often trips up the uninitiated (or the sleepy). A more rigorous unit would be nH/turns^2, which uses the same numerical value as A_L.

In addition to the inductance value and parasitic shunt capacitance, effective series resistance is of great importance. To estimate it, one would be tempted quite naturally to make use of the skin effect formula. Unfortunately, that formula assumes a uniformly illuminated semi-infinite block of conductor. In a solenoid, however, the conditions are quite

11. Medhurst claims much better accuracy for his formula, but in fact his maximum error is as large as 8%.

different: the magnetic field of one turn affects the current distribution of neighboring turns, so that the boundary conditions (and consequently, the effective cross-sectional areas) are considerably modified. Use of the skin effect formula therefore usually leads to rather gross errors. In particular, it predicts that *Q* should ultimately increase as the square-root of frequency, since the inductive reactance increases linearly with frequency, while the skin resistance grows as its square-root. In reality, *Q* does increase approximately in this fashion only at the lower frequencies, then generally reaches a roughly constant value over a reasonably broad frequency range, before plummeting as resonance is approached. The broad constant range is due to eddy current losses in one turn induced by the current flowing in nearby turns. These losses increase approximately linearly with frequency, causing the ratio of inductive reactance to effective resistance to approach a constant value[12]. Then, as one approaches resonance, the net reactance plunges, causing *Q* to do so as well.

Now, we know that interaction among turns must be considerable, for if it were not, inductance would grow only linearly, rather than quadratically, with the number of turns. It is this interaction that also explains the loss behavior, as well as why the maximum *Q* occurs for a particular turn-to-turn spacing: If the turns are too close together, the interaction greatly reduces the effective cross-sectional area, thereby increasing resistance. If the turns are too far apart, the total wire length increases, again increasing total winding resistance. At some intermediate value of spacing, *Q* is maximized. A number of studies have shown that using a winding pitch approximately equal to twice the wire diameter, setting the solenoid length at approximately twice the diameter, and using the largest practical diameter allowed by, say, self-resonance criteria, produces the maximum *Q* for typical solenoids[13]. Fortunately, the optimum conditions are relatively flat, so the *Q* value achieved is not overly sensitive to departures from the optimum conditions.

Given the foregoing observations, one may apparently use a skin-effect based calculation only to establish a lower bound on the series resistance. Sadly, there is no simple general formula to predict the effective RF resistance for an arbitrary coil design, despite a considerable number of efforts dating back to the 1920s. Readers are invited, even encouraged, to take up this problem and solve it. The best that exists presently is still perhaps the complicated formula, involving multiple lookup tables, found in Terman's classic, *Radio Engineer's Handbook.* [14]

The last bit of data that might be useful in designing these coils concerns the properties of wire. The conductivity of pure copper is about 5.7×10^7 S/m. The diameter of bare copper wire is usually presented in tabular form, but a simple (though approximate) formula is

12. Dielectric loss behaves similarly, and may contribute significantly to *Q* degradation if the electric field of the inductor permeates lossy dielectric materials.

13. For a review of some of these studies, see *RCA Radiotron Handbook*, 2nd ed., 1942. See also F.E. Terman, *Radio Engineer's Handbook*, first edition, McGraw-Hill, 1943. Terman shows that the optimum pitch is actually a weak function of the length-to-diameter ratio. Since the optimum *Q* is itself not a strong function of pitch, the rule of thumb given is usually adequate.

14. McGraw-Hill, first edition, 1943, pp. 77-83. The equations presented there are based largely on the extensive work of Butterworth in the late 1920s and early 1930s.

$$D \approx \frac{0.32}{10^{(AWG)/20}}, \tag{14}$$

where the diameter *D* is in *inches*, and *AWG* is the (American) wire gauge. This formula yields values correct to within about 2% for wires between 10 and 40 gauge, a range that spans the most commonly used sizes. Note that it implies a decrease in diameter by a factor of ten for every wire gauge increase of 20, so the relative behavior of the wire gauge on diameter is the same as that of voltage expressed in dB.

There is no correspondingly simple formula for enameled wire, but adding an arbitrary 0.0045" to the values for bare wire yields diameters that are typically correct to approximately 5% or better. It should be mentioned that insulator thicknesses vary somewhat from manufacturer to manufacturer, so values calculated from these equations must be verified in all cases where it matters. These formulas are presented mainly as guides for back-of-the-envelope types of calculations.

4.4 Single loop

Another useful formula is for the inductance of a single loop. Despite the simplicity of the structure, there is no exact, closed-form expression for its inductance (elliptic functions arise in the computation of the total flux). However, a useful "cocktail napkin" approximation is given by:

$$L \approx \mu_o \pi r. \tag{15}$$

This formula tells us that a loop of 1mm radius has an inductance of approximately 4nH.

In deriving this approximation, the flux density in the center of the loop is arbitrarily assumed to be one-half the average value in the plane of the loop, then the inductance is computed as simply the ratio of total flux to the current. In view of the rather coarse approximation involved, it is remarkable that the formula does as well as it typically does.

Note that, for a single turn and in the limit of zero length, Wheeler's formula (Eqn. 8 and Eqn. 11) converges to within about 11% of $\mu_o \pi r$.

Much better accuracy is provided by the following equation, which takes into account a nonzero wire diameter as well as magnetic coupling among infinitesimal wire segments[15]:

$$L \approx \mu_o r \left[ln\left(\frac{8r}{a}\right) - 2 \right], \tag{16}$$

where *a* is the radius of the wire. With this equation, we see that Eqn. 14 strictly holds only for an *r*/*a* ratio of about 20.

15. Ramo, Whinnery and Van Duzer, *Fields and Waves in Modern Radio*, Wiley, 1965, p. 311.

To make a crude approximation even more so, Eqn. 14 can be extended to non-circular cases by arguing that all loops with equal area have about the same inductance, regardless of shape. Thus, we may also write:

$$L \approx \mu_o \sqrt{\pi A}, \tag{17}$$

where *A* is the area of the loop. A closed contour of one square centimeter area has an inductance of about 7nH according to this formula. This equation, very approximate as it is, turns out to be quite handy in estimating the magnitude of various component and lay-out parasitics, as well as in evaluating the likely efficacy of proposed layout changes.

We can check the reasonableness of these equations by considering the inductance of a loop of extremely large radius. Since we can treat any suitably short segment of such a loop as if it were straight, we can use the equation for the inductance of a loop to estimate the inductance of a straight piece of wire.

We've already computed that a circular loop of 1mm radius has an inductance of 4nH, so we have roughly 4nH per 6.3mm length (circumference), which is in the same range as the value given by the more accurate formulas.

4.5 Magnetically coupled conductors

The magnetic fields surrounding conductors drop off relatively slowly with distance. As a result, there can be substantial magnetic coupling between adjacent (and even more remote) conductors. A measure of this coupling is the *mutual* inductance between them. For two infinitesimally thin round wires of equal length, this inductance is given approximately by:

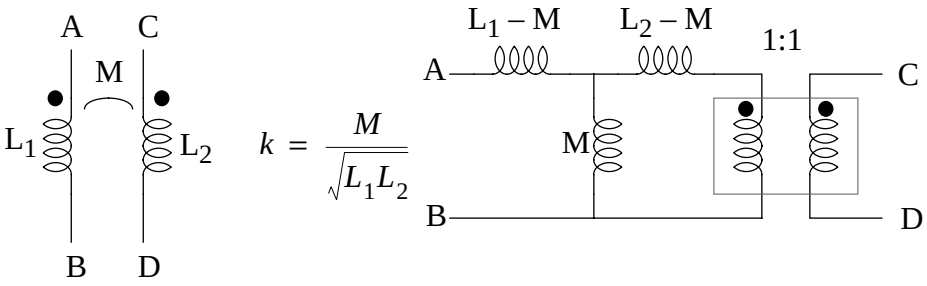
$$M \approx \frac{\mu_o l}{2\pi}\left[ln\left(\frac{2l}{D}\right) - 1 + \frac{D}{l}\right], \tag{18}$$

where *l* is the length of the wires, and *D* is the distance between them.[16] For a 10mm length and a spacing of 1mm, the mutual inductance works out to about 4nH. Since the inductance of each wire in isolation is about 10nH, the 4nH mutual inductance represents a coupling coefficient of 40%. The logarithmic dependence of *M* on spacing means that the coupling decreases rather slowly with distance, so one must be aware of the possibility of unwanted coupling between non-adjacent conductors.

One model for coupled inductors is an inductive T-network in cascade with an ideal transformer:

16. This formula is adapted from Terman, op. cit.

FIGURE 4. Coupled inductors and circuit model



In this model, L_1 and L_2 are the values each inductor has with no current flowing in the other. Resistive losses, as well as parasitic capacitances, are not shown in the ideal model of Figure 4, but should be taken into account in critical designs.

5.0 Summary

We've seen that seemingly ordinary components must be modeled in progressively more sophisticated ways as frequency increases. Nominally simple components are seen to have important behaviors that may be ignored only at low frequencies. Even resistors, capacitors and inductors must be treated as complicated impedances for proper design of microwave circuits. As an aid to developing appropriate models, this chapter has presented numerous equations and rules of thumb for estimating parasitic inductance and capacitance.

Time Domain Reflectometry

1.0 Introduction

Both time- and frequency-domain characterizations provide comprehensive information about a system. The latter require the ability to generate and measure sinusoidal voltages and currents over a broad frequency range. The network analyzer, in either scalar or vector incarnations, is an example of such an instrument.

An alternative is to use time-domain methods to characterize a system. The principal tool of this type is the time-domain reflectometer (TDR), which is in essence a miniature radar system. The TDR launches a pulse (“the main bang”) into the device under test, and observes any echoes. The timing of a reflection with respect to the main bang indicates the location of a discontinuity, and the shape of the reflected pulse conveys important information about its nature. With a reflectometer, then, one can quickly locate and characterize both resistive and reactive discontinuities (and evaluate their fixes). A network analyzer can provide this information as well, but requires considerably more labor to do so.

In this chapter, we study how a TDR works, and then show how to build a simple pulse generator that is a handy instrument in its own right. A small modification of the pulse generator yields a fast risetime step generator which can be used in conjunction with a fast oscilloscope to construct a surprisingly inexpensive TDR with sub-nanosecond capability.

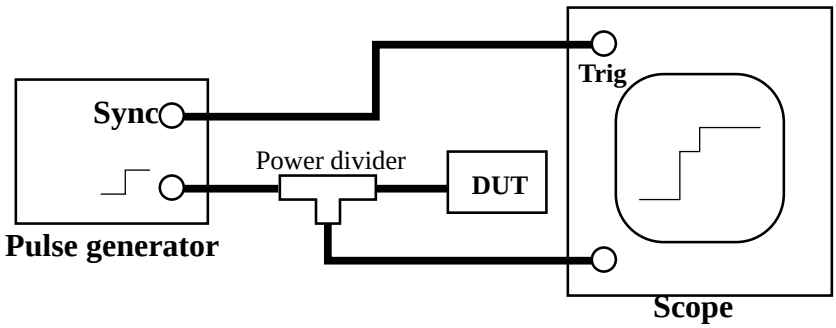
2.0 Applications of TDRs

There are two primary applications of TDRs: finding and characterizing impedance discontinuities. These capabilities translate directly into the ability to correct defects and evaluate the quality of any compensation performed.

2.1 Locating discontinuities

A TDR consists of just two main modules: a pulse generator, and an oscilloscope:

FIGURE 1. Time domain reflectometer



A pulse generator applies a fast risetime step to the device under test (DUT). A portion of the signal is tapped off and fed to an oscilloscope, whose sweep is synchronized with the step. The synchronizing signal is timed to allow the display of the voltage both a bit before the rising edge and well after.

A pulse's risetime determines its spectral content and, hence, the bandwidth over which the TDR can perform a useful characterization. Similarly, the oscilloscope's bandwidth must be consistent with the desired characterization bandwidth. A common rule of thumb is that the –3dB bandwidth of a step is related inversely to the 10-90% risetime as follows:

$$f_{-3dB}t_r \approx 0.35 \tag{1}$$

This relationship, although strictly correct only for single-pole systems, allows us to estimate the performance requirements of a TDR system. For example, suppose we wish to characterize a transmission line up to 10GHz. Using our rule of thumb, we find that the TDR's risetime must be shorter than about 35ps. The fastest commercially available TDRs are capable of characterizing systems beyond 50GHz, implying risetimes of under 7ps.[1]

The risetime of the incident pulse determines not only the bandwidth over which the system is characterized, but also the spatial resolution of the characterization. If a pulse reflects off of a discontinuity some distance x_d from the source, the total time taken in the round trip back to the source is

$$t_{prop} = \frac{2x_d}{v_{prop}} \tag{2}$$

so that

$$x_d = \frac{t_{prop}v_{prop}}{2} \tag{3}$$

where v_{prop} is the propagation velocity. Clearly, if the pulse's risetime is too slow, then reflections will be obscured during the rising edge. Roughly speaking, the spatial resolution is approximately equal to the distance traveled during the risetime. The first reflectometers were developed to locate faults in very long cables, where the ability to pin down the location of an open or short to within 100 meters or so suffices. Given that the speed of light along a typical cable is about 60-80% of the free space value, the corresponding delay is about 4ns per meter. Risetimes in the range of hundreds of nanoseconds, implying bandwidths in the low MHz range, therefore can be satisfactory for such cable fault-finding applications. The far faster sub-10ps risetimes cited earlier for today's leading edge gear correspond to the ability to locate discontinuities to a resolution of a few millimeters in free space. Such risetimes and their corresponding spatial resolutions are much more compatible with the size of typical microwave circuit elements and modules.

1. Here we are excluding systems that employ cryogenics and superconductors.

It is not necessary to know the propagation velocity to locate a discontinuity, despite the seeming implications of Eqn. 3. With microstrip, for example, just run a finger along the line while observing the TDR trace. When the bump produced by your finger coincides with the bump produced by the discontinuity you're trying to investigate, you've found it: the discontinuity will be right underneath your finger. (Of course this method should not be used if the TDR pulse is of an unusually high power!)

2.2 Characterizing discontinuities

One reason that the TDR is so valuable is that it conveys much more information than merely the location of discontinuities. That this is so is most directly understood from the relationship between the reflection coefficient and the termination impedance:

$$\Gamma = \frac{Z_{Ln} - 1}{Z_{Ln} + 1} \tag{4}$$

where Z_{Ln} is the normalized load impedance:

$$Z_{Ln} \equiv \frac{Z_L}{Z_0} \tag{5}$$

Note that the reflection coefficient is a complex quantity in general, possessing both a magnitude and phase (or real and imaginary part). It thus contains information about how the spectral components of the step response are modified in reflecting off of the discontinuity. Note also that Γ contains similarly complete information about the load impedance (Eqn. 4 may be solved for Z_{Ln} in terms of Γ). Adding the assumption of linearity allows us to bring to bear on the problem all of the powerful tools of linear system theory. In particular, finding the step response is a moldy staple of system theory, and that is precisely what the TDR displays. Even though we'll start with a formal mathematical approach, we'll quickly examine a few representative cases to extract physical insight to see how one might guess the correct answer for these and many other cases of practical relevance.

The response to any input is the sum of the input excitation as well as any reflection that arises. The reflection is merely Γ times the incident signal. Hence the transfer function that relates the total output to the input is

$$H(s) = 1 + \Gamma = \left(\frac{2Z_{Ln}}{Z_{Ln} + 1} \right) \tag{6}$$

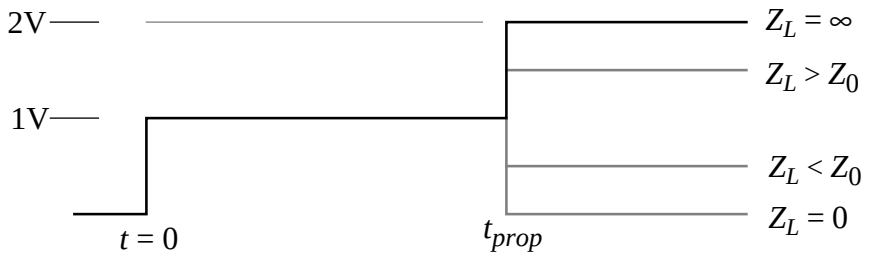
When using this equation, it's important to keep track of the fact that the inverse Laplace transform of Eqn. 6 is only valid for times greater than the roundtrip time-of-flight,

$$t_{prop} = \frac{2x_d}{v_{prop}} . \tag{7}$$

Before this time, the response is just the value of the input alone (e.g., one volt if we have assumed a unit step excitation).

Using these relationships, it is straightforward to determine the TDR traces for several commonly encountered cases. For example, consider open- and short-circuit loads. In those two cases, the normalized load impedances are infinite and zero, respectively, with corresponding values of two and zero for $H(s)$. Keeping in mind that these values apply only after the time-of-flight delay, the unit step responses thus appear as follows:

FIGURE 2. Idealized TDR trace for open, shorted and resistive loads



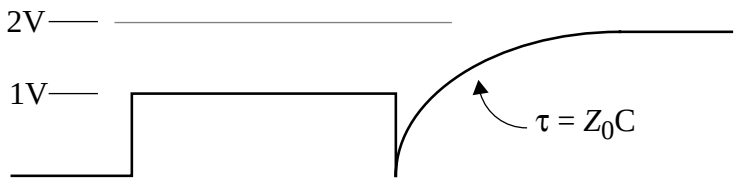
For resistive loads in between these two extremes, the step response will jump to some level between zero and 2V. If the load resistance is less than the characteristic impedance, the final value will be below 1V. If greater, the final value will lie between 1V and 2V; and if equal to Z_0, no discontinuity will be observed.

Now consider the step response when reactive loads terminate a line. If the load element is a capacitance, then

$$\frac{2Z_{Ln}}{Z_{Ln}+1} = \frac{2}{1+\dfrac{1}{Z_{Ln}}} = \frac{2}{1+sZ_0C} \tag{8}$$

This is simply the transfer function of a single-pole low-pass filter, whose step response should be familiar:

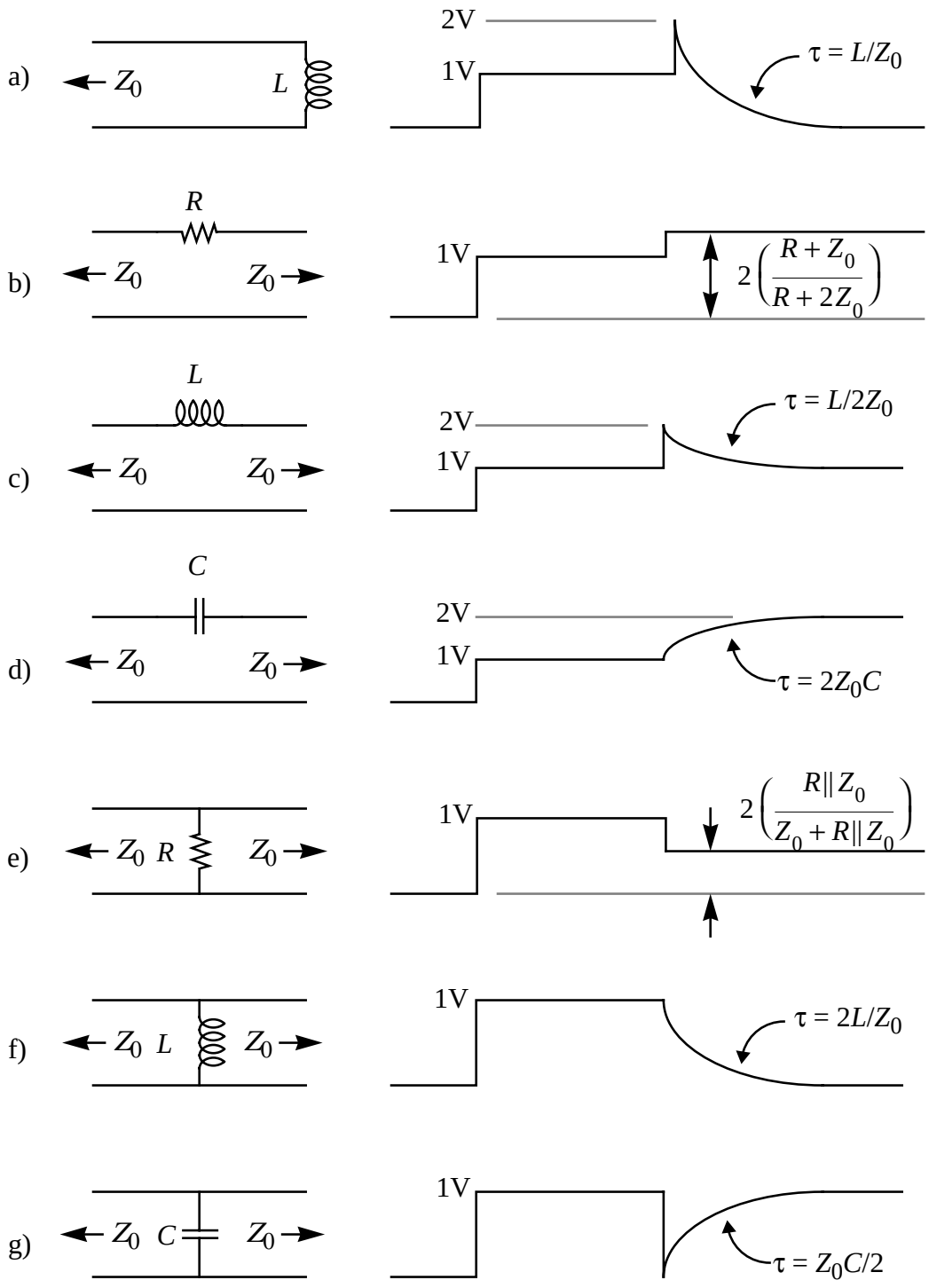
FIGURE 3. Idealized TDR trace of capacitively terminated transmission line



In like manner, the step response for any number of discontinuities can be readily determined. Without providing detailed derivations (which are left as a pleasant exercise for the

reader), here is a short catalog of simple, but practically relevant, discontinuities and their corresponding TDR traces:

FIGURE 4. Idealized TDR traces for several discontinuities (incident amplitude = 1V in all cases)



The shapes of the TDR traces can be anticipated from purely physical arguments with a minimum of mathematics. In all of the reactive examples, there is only one time constant because we have considered only single-reactance loads. A single time constant implies a single exponential factor. An inductive termination (case a) appears initially as an open circuit, but ultimately acts as a short. The time constant of the exponential transition between these two conditions is the ratio of inductance to the effective resistance it sees (here, Z_0). In case c, the inductance sees a total resistance of $2Z_0$ (one Z_0 each to the left and right), and the final value is 1V.

In case b, that of a series resistive discontinuity, the step response must jump up because the effective load resistance is the resistance as viewed from the discontinuity to the right. Here, that is the sum of R and Z_0. A simple voltage divider equation yields the result shown in the figure (just remember that the open-circuit step amplitude is 2V).

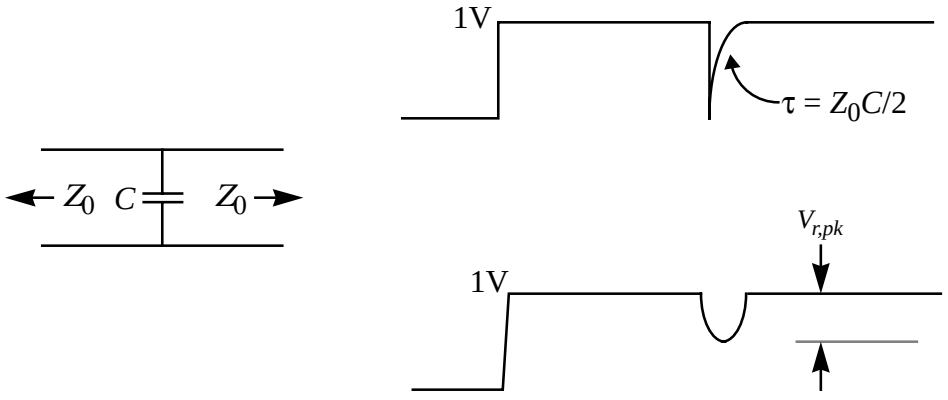
Arguments similar to the foregoing can be used to sketch the TDR traces for the rest of the examples given.

In practice, the observed TDR traces will differ (perhaps greatly) from the idealized ones shown in the figure. The main difference is due to the finite risetime of the step excitation. If one considers a practical step to be the result of low-pass filtering an infinitely fast one, the actual TDR traces may be deduced by low-pass filtering the ideal traces through a filter whose step response has the same risetime as that of the actual step. This filter will slow down rising edges and cause a rounding of sharp corners.

2.3 Parameter extraction

Using our catalog of TDR traces, it is often possible to measure small inductances and capacitances, or even extract a more complex circuit model, from a measured step response. To do so requires that we consider explicitly how the limited bandwidth of all real systems affects the shape of the waveform. As a specific example, consider a shunt capacitive discontinuity:

FIGURE 5. Ideal and more realistic TDR traces for shunt capacitance



The reflection coefficient is

$$\Gamma = \frac{Z_{Ln} - 1}{Z_{Ln} + 1} = \frac{\dfrac{1}{sCZ_0 + 1} - 1}{\dfrac{1}{sCZ_0 + 1} + 1} = \frac{-sCZ_0}{sCZ_0 + 2} \tag{9}$$

The reflection coefficient is seen to have a pole at a frequency given by

$$\omega = \frac{2}{CZ_0} \tag{10}$$

Spectral components above this pole frequency are attenuated by the low-pass filter effectively formed with the capacitor. This filtering is the reason for the change in shape shown in Figure 5. The sharp edge gets smeared out, resulting in the smooth bump shown in the bottom trace. If, as is usually the case, the capacitive discontinuity is small enough that this filtering effect may be neglected, then the reflection coefficient may be approximated by

$$\Gamma \approx \frac{-sCZ_0}{2} \tag{11}$$

The incident and reflected signals thus may be related approximately by a derivative:

$$V_r = \Gamma V_i \approx \frac{-CZ_0}{2}\frac{dV_i}{dt} \tag{12}$$

The peak value of V_r is proportional to the peak value of the input slope, and thus C is approximately

$$C \approx \frac{-2V_r}{Z_0}\left(\frac{dV_i}{dt}\right)^{-1} \approx \frac{-2V_{r,\,pk}}{Z_0}\left(\frac{V_i}{\tau}\right)^{-1} \tag{13}$$

where $V_{r,pk}$ is as shown in Figure 5, V_i is the amplitude of the input step, and τ is the time constant of the input step. (We have used the fact that, for a single-pole system, the maximum slope of the step response is simply the amplitude of the step, divided by the time constant.) The 10%-90% risetime of a step is approximately equal to 2.2τ, so we could also write:

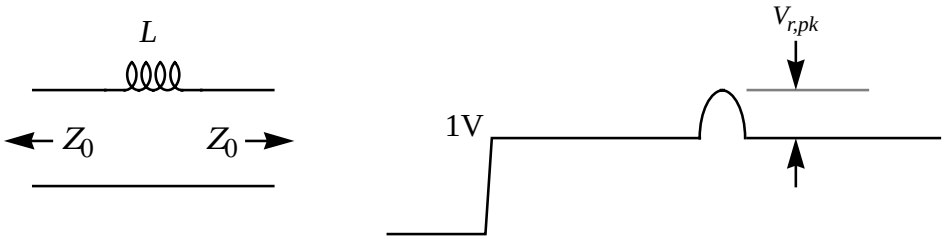
$$C \approx \frac{\left|V_{r,\,pk}\right|}{1.1Z_0}\left(\frac{t_{rise}}{V_i}\right). \tag{14}$$

An analogous derivation for the case of a series inductive discontinuity yields the following estimate:

$$L \approx 2Z_0 V_r \left(\frac{dV_i}{dt}\right)^{-1} \approx \frac{Z_0 V_{r,\,pk}}{1.1}\left(\frac{t_{rise}}{V_i}\right), \tag{15}$$

where a typical waveform is as follows:

FIGURE 6. Somewhat more realistic TDR trace for series inductance

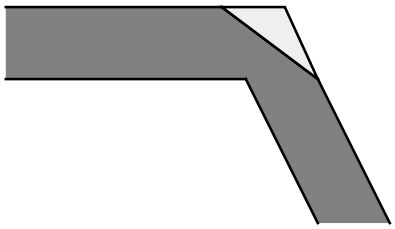


This method can provide remarkable measurement resolution. Suppose, for example, that a given TDR system possessed the ability to resolve a voltage as small as 1mV, along with a 15ps risetime, 1V step. The smallest measurable capacitance and inductance would be about 0.3fF and 0.7pH! Needless to say, it is exceedingly difficult to make measurements of such small values using any other method. From this calculation, it is clear that even relatively insensitive, slow TDR systems are capable of impressive measurements of inductance and capacitance.

2.4 Compensation

By identifying the location and type of discontinuity, the TDR enables you to design compensators, should that prove necessary. Because of the speed with which TDR characterizations may be performed, the efficacy of any compensation scheme is rapidly evaluated. As a specific example, consider the mitered bend:

FIGURE 7. Mitered bend



The optimum amount of mitering is easily determined experimentally with a TDR. Pieces of the corner are sliced off until the reflections are minimized. To achieve this same result with, say, a vector network analyzer, or a slotted line SWR measurement would require more (and perhaps considerably more) work.

An important consideration is that a given discontinuity may mask the existence or size of other discontinuities further down the line. For example, a large series inductance (or a

large shunt capacitance) may reduce the bandwidth of the TDR pulse downstream of the discontinuity, reducing the ability to characterize discontinuities past the inductor. Therefore, the proper method is to fix the discontinuity nearest the source first, retest with the TDR, fix the next discontinuity, and so forth until all problems are fixed.

3.0 Summary and Conclusions

The TDR is an indispensable complement to traditional frequency-domain equipment, permitting the characterization of microwave systems over a broad frequency range in a remarkably short time. The ability to locate discontinuities is a particularly valuable capability of TDRs, as is the related ability to evaluate expediently the quality of any compensation methods over a broad band of frequencies.

4.0 Projects: Homegrown Fast Pulse and Step Generators

The art of fast pulse and step generation is highly specialized, and it is unrealistic to expect to generate pulses with risetimes competitive with state of the art instruments using what's available in the typical home laboratory. However, you may be pleasantly surprised to find that it isn't difficult to generate pulses with risetimes in the neighborhood of 200ps using components readily available to hobbyists. Such a pulse generator is especially valuable for evaluating the quality of oscilloscopes and particularly of scope probes.

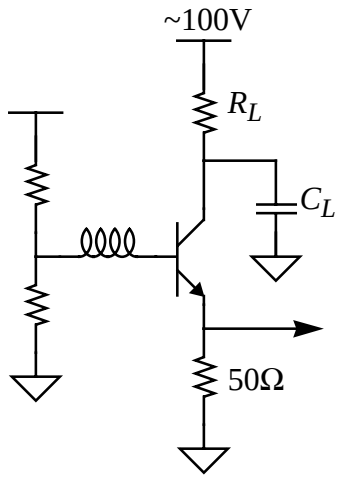
A trivial modification to the pulse generator converts it into a triggerable step generator with 200ps risetime. The risetime is short enough to locate discontinuities to a resolution of approximately 7cm in free space, or roughly 5cm in microstrip. When coupled with a suitably fast oscilloscope, the step generator enables the time-domain characterization of circuits up to approximately 2GHz, making it a good match with the homegrown focus of this book, with BNC connectors mated to microstrip on FR4.

4.1 Free-running sub-ns pulse generator

Of the possible ways to generate fast pulses, the most economical for hobbyists is unquestionably to make use of an abnormal mode of transistor operation: *avalanche breakdown*. In this type of breakdown, the collector voltage is high enough to rip electrons from their orbits, creating hole-electron pairs. The electrons accelerate toward the positive terminal (here assumed to be the collector), while the holes accelerate toward the base. As the freed carriers accelerate, some bash into other silicon atoms, creating still more hole-electron pairs, and so on, causing a rapid increase in collector current.

The following pulse generator circuit exploits this avalanching:

FIGURE 8. Avalanche mode pulse generator



In this circuit, the collector supply voltage is chosen well above the transistor's breakdown voltage, and its precise value is not at all critical. **However, under no circumstances should you derive this voltage directly from the mains**; it is simply too dangerous to do so! Rather, a battery-operated circuit is highly recommended. A particularly handy source of high voltage is the xenon flash circuitry of disposable cameras. These often may be obtained at low cost (or even free) from neighborhood photo labs. Typically 200-300V may be found across the one large capacitor in those units, and this level of voltage is more than adequate to avalanche almost any transistor of interest. Exercise caution when removing the board from the camera case, and certainly while wiring it up. The main capacitor can store a dangerous amount of charge for quite some time.

The capacitor C_L may be made out of copper foil tape over ground plane on 1/16" FR4. A good starting value is a square strip approximately 0.75cm on a side. Foil may be added or trimmed as necessary to adjust pulse duration and amplitude.

Once avalanching begins, the collector current increases rapidly for two reasons. One is the direct effect of avalanche electron multiplication in the collector, and the other is the increase in base current produced by the avalanching holes. The increased base current increases the collector current through ordinary transistor action. This positive feedback mechanism is enhanced by biasing the base through a relatively large impedance to allow the hole current that comes *out* of the base to raise the base voltage significantly.

The collector load resistor R_L is quite large, so the collector current is actually supplied by the capacitor during the avalanche time. The large collector current quickly depletes the charge in the capacitor C_L, dropping the collector voltage below the avalanche threshold, rapidly terminating the flow of current. The capacitor recharges slowly through the R_L, and eventually causes another avalanche. The pulse repetition frequency is therefore determined by the product of R_L and C_L, and is typically in the range of tens to hundreds of kHz with commonly used values.

The pulse width depends on the size of the collector capacitor (larger capacitances lead to taller and wider pulses) and the characteristics of the transistor. Low collector-base capac-

itance is favored to allow the base and collector to move in opposite directions. A more critical parameter is the ratio of BV_{CBO} to BV_{CEO}. The former is a measure of collector-base breakdown voltage with the emitter open-circuited, while the latter is the breakdown voltage with the base open-circuited. The latter is always smaller than the former precisely because of the same positive feedback mechanism already described. For most small-signal transistors, the ratio of these two breakdown voltages falls within the range 1.5-2, but a few (such as the 2N2369 or Zetex FMMT-417) exceed 2.5 or so. Those few are the ones that are particularly well suited for making avalanche pulsers.

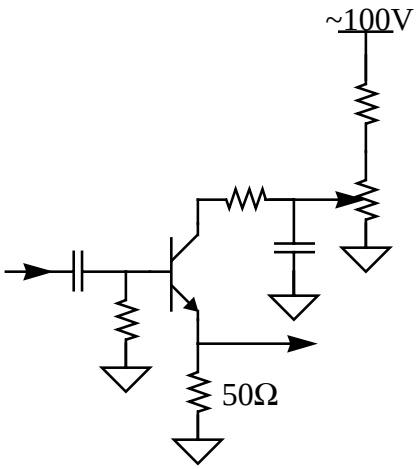
It is important to underscore that transistors are almost never specified by manufacturers for avalanche mode operation. Even if you do find a transistor that avalanches well, you should not expect all transistors of a given type to avalanche similarly. Consequently some hand selection will generally be necessary. That said, typically more than 75% of a given batch of 2N2369 transistors will avalanche well enough to provide a 5-10V peak pulse into 50Ω with rise and fall times close to 200ps, speeds which were state of the art for expensive laboratory instruments in the mid-1960s. The ~1A/ns current slew rate is quite difficult to achieve through conventional means, so having to try a handful of transistors seems a modest price to pay indeed. This high a slew rate also underscores the importance of assiduously reducing parasitic inductance in series with the emitter circuit: just 1nH of stray inductance drops a volt!

The pulse generator is a versatile instrument with multiple uses in high-speed work. As one specific example, the bandwidth of a scope-probe combination can be rapidly evaluated with such a generator by observing the displayed rise and fall times. Aberrations introduced by defective or improperly calibrated instruments and cables are also readily observed. Given that the typical alternative is to measure frequency response by sweeping a sinewave generator over a GHz range, the pulse generator is clearly an extremely inexpensive option.

4.2 Triggerable sub-ns step generator

Sometimes it is more convenient to generate fast pulses that possess widths significantly longer than the risetime, that is, we might wish to generate approximations to step waveforms. This is particularly so for TDR purposes, where we often desire to evaluate step responses directly. Fortunately, we can modify the pulse generator without too much trouble and convert it into a step generator:

FIGURE 9. Avalanche mode step generator



Here, the potentiometer is adjusted to place the transistor just below the threshold of spontaneous avalanching. A positive-going trigger pulse pushes the transistor over the edge, initiating the avalanche. The output pulse duration is a function of the size of the collector capacitance. For best results, the collector capacitance should be realized as the parallel combination of a microstrip and a chip capacitor. The pulse width should be chosen much longer than the full roundtrip time along the longest line you desire to characterize. The risetime determines the shortest distance away from the source where a discontinuity may be discerned. The shorter the risetime, the shorter the resolvable distance.

This circuit is a bit twitchy, and periodic adjustment of the potentiometer is generally necessary to compensate for drift with temperature, and for changes in the actual value of the high voltage.

The need for a trigger can be removed by adjusting the potentiometer well above the avalanche voltage. The circuit then free runs with a pulse repetition frequency determined by the collector load network. The circuit remains sensitive to the trigger, and synchronization of the circuit to an external signal whose frequency is somewhat above the free-running frequency is possible.

5.0 Further Reading

An excellent applications note on the use of the TDR may be found in the February, 1964 issue of the Hewlett-Packard Journal (vol. 15, no. 6).

Network Analyzers

1.0 Introduction

The development of the automatic vector network analyzer (VNA, or simply *network analyzer*) has revolutionized the characterization of microwave circuits. By computing all of the s-parameters of a network over a broad frequency range, the network analyzer provides the designer with a comprehensive overview of circuit behavior that would be extremely cumbersome to obtain manually.

This chapter begins with a description of one instrument that the VNA has largely displaced, the slotted line. There are several motivations for this retrospection. One is pedagogical, for the slotted line affords us an opportunity to investigate directly the quintessential wave phenomena of reflection and interference. Another is practical, because the slotted line exploits these phenomena to measure impedance at high frequencies with comparatively inexpensive equipment. Yet another is that important calibration issues that also apply to the VNA are quite naturally introduced with the slotted line. Finally, the labor involved in making accurate measurements with a slotted line is large enough to explain what motivated development of the network analyzer.

A detailed description of the VNA follows, along with illustrative examples of how it is used. There is a focus on identifying and mitigating sources of error, along with a comprehensive description of calibration techniques, because much of the modern VNA's power derives from its ability to characterize and remove its own errors.

The chapter concludes with instructions on how to build an inexpensive slotted line system capable of measuring impedance over a 1-5GHz range.

2.0 The Old Days: Slotted Line Impedance Measurement

Prior to the development of the network analyzer, characterization of microwave systems was a cumbersome process. Consider first the basic problem of measuring impedance. At low frequencies, it is a relatively simple matter to use a bridge measurement technique, or to excite a network with, say, a voltage and measure the current that flows in response. Finding the ratio of voltage to current is straightforward, even if one must keep track of the relative phase between them in order to compute both the real and imaginary parts of the impedance:

$$Z = \frac{V}{I} = |Z| e^{j\phi} \tag{1}$$

As frequency increases, however, the situation gets progressively more complicated. Adding to the usual difficulties associated with making instruments operate at high frequencies

are the very significant problems of fixturing: the impedance of a given length of conductor perturbs the measurement more and more significantly as frequency increases.

A recurring theme in good engineering is the conversion of a liability into an asset ("it's not a bug, it's a *feature*"). In this case we acknowledge *a priori* the futility of trying to reduce fixturing impedances to insignificant levels. Rather than attempt to quantify and remove the effect of the fixturing on the measured impedance, we consider instead the effect of the load impedance on the fixturing. To understand why this change in viewpoint is so valuable, recall that voltage is independent of position only along a properly terminated transmission line (or waveguide). Any mistermination gives rise to a reflection which periodically interferes constructively and destructively with the incident wave, producing standing waves along the line. The amplitude and phase of the standing waves depend uniquely on the mismatch between the load impedance and the line's characteristic impedance, Z_0. Measurement of the standing waves, coupled with knowledge of Z_0, thus allows computation of the load impedance Z_L. The core of this impedance measurement method is therefore the bi-unique relationship between impedance and reflection coefficient:

$$Z_{Ln} = \frac{1+\Gamma}{1-\Gamma} \tag{2}$$

where Z_{Ln} is the normalized load impedance,

$$Z_{Ln} \equiv \frac{Z_L}{Z_0}, \tag{3}$$

and the reflection coefficient is a complex quantity:

$$\Gamma = |\Gamma| e^{j\phi} \tag{4}$$

The mathematical basis for the measurement technique becomes clear by first expressing the voltage along a transmission line as the sum of forward and reflected components:

$$V(z) = V_f + V_r = V_f(e^{-j\beta z} + \Gamma e^{j\beta z}) = V_f e^{-j\beta z}(1 + \Gamma e^{j2\beta z}) \tag{5}$$

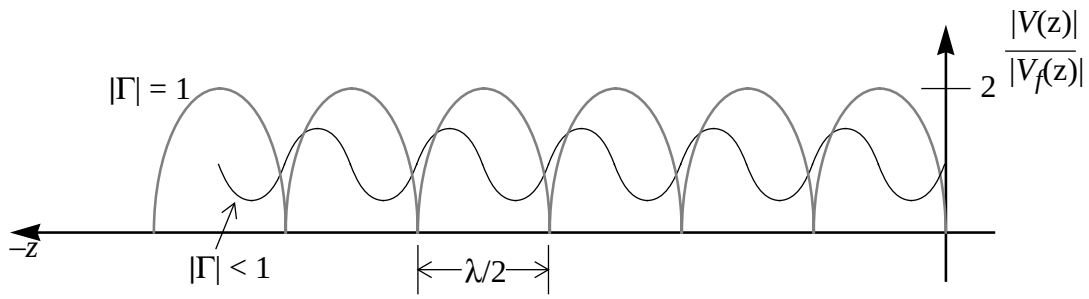
where $z = 0$ is defined as the location of the load, with z increasingly negative as one approaches the source, and β is the phase constant, $2\pi/\lambda$.

The magnitude of the line voltage as a function of position is

$$|V(z)| = \left|V_f e^{-j\beta z}\right| \left|(1 + \Gamma e^{j2\beta z})\right| = \left|V_f\right| \left|(1 + \Gamma e^{j2\beta z})\right| = \left|V_f\right| \left|(1 + |\Gamma| e^{j(\phi + 2\beta z)})\right| \tag{6}$$

where ϕ is the phase angle of the reflection coefficient. Note from Eqn. 6 that the voltage magnitude is periodic. These standing waves have a periodicity of $\lambda/2$, so the distance between, say, minima corresponds to π radians of phase:

FIGURE 1. Typical plot of amplitude vs. position for two values of Γ



The minimum and maximum voltages occur when the exponential factor is –1 and +1:

$$V_{min} = |V_f|\,(1 - |\Gamma|) \tag{7}$$

$$V_{max} = |V_f|\,(1 + |\Gamma|) \tag{8}$$

The *standing wave ratio* (SWR) is defined as the ratio of maximum to minimum voltage:

$$\text{SWR} = \frac{V_{max}}{V_{min}} = \frac{1 + |\Gamma|}{1 - |\Gamma|} \tag{9}$$

From Eqn. 9 it is clear that measurement of SWR allows the computation of |Γ|.

To complete the measurement, we need ϕ, the phase of Γ. The key is to note that the minimum voltage occurs when

$$1 + |\Gamma|\,e^{j(\phi + 2\beta z)} = 1 - |\Gamma| \tag{10}$$

or, equivalently,

$$\phi + 2\beta z = (2n + 1)\,\pi \tag{11}$$

where *n* is any integer. Therefore, the phase of Γ can be computed from:

$$\phi = (2n + 1)\,\pi - 2\beta z \tag{12}$$

where *z* is the location of the minimum (again, *z* is a negative quantity in our coordinate system).[1]

A practical consideration is that the precise location of the *electrical* reference plane *z* = 0 is not always obvious. As a consequence an experiment is generally required to determine this piece of information. The traditional (and simplest) method is simply to terminate the line in as good a short circuit as possible. Clearly, the minima will be nulls (ideally, any-

1. In principle, one could also use the maxima in the measurement. However, the minima are sharper, so that a given amplitude measurement uncertainty translates into a smaller (perhaps much smaller) timing uncertainty than if the maxima were used.

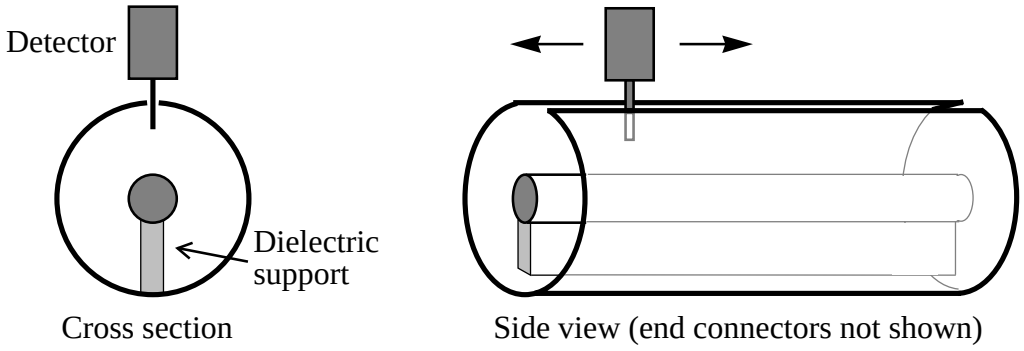
way), again periodically disposed along the line. Any of these nulls may be taken as the location of the reference plane $z = 0$, although it is customary to choose the one closest to the short. Pick one and record its position. Also note the spacing between successive minima (this is equal to $\lambda/2$), so that you can readily compute the phase constant β. Then replace the short with the impedance to be measured. Measure SWR to enable a calculation of $|\Gamma|$, and note the shift in the position of the minima relative to the zero reference established with the shorted load, counting shifts away from the load as having a negative sign. Plug that value into Eqn. 12 and solve for ϕ. Then use Eqn. 4 in Eqn. 2 to find the (normalized) load impedance. The actual load impedance is found simply by multiplying this value by Z_0.

The beauty of this technique is that the fixturing does not need to be short compared with a wavelength. In fact, the fixture's length actually must exceed a half wavelength (and preferably be several half wavelengths) in order for standing waves to be characterized.

The measurement requires knowledge of the voltage as it varies along the line. In turn this requires that we have physical access to the line. A slotted-line system therefore consists of an air-dielectric transmission line (or waveguide) that is slit open to admit a probe (which is generally a simple high impedance diode detector capacitively coupled to the line). The slit and probe are carefully designed to minimize disturbance of the fields. In the case of a coaxial line, a lengthwise slit in the outer conductor has a minimal effect because no currents flow circumferentially. The primary effect of the slit is a small reduction in capacitance per unit length and a consequent small increase in characteristic impedance. A suitably narrow slit minimizes this effect to negligible levels and also ensures a minimum of radiation and its attendant losses.[2]

The probe is mounted on a slider with a calibrated ruler so that its position along the line can be measured (a coaxial line is shown, but a slotted waveguide also works):

FIGURE 2. Coaxial slotted line



In most slotted lines, the probe's depth into the line is adjustable, allowing a tradeoff between detector sensitivity and disturbance of the field pattern. Fortunately, the probe's

2. As long as the slit is comparable to, or narrower than, the wall thickness, it will act much like a waveguide far beyond cutoff. As a result, it is almost always the case that radiation can be considered truly negligible, and many texts don't even bother to mention the possibility of radiation at all.

presence does not affect the location of nulls (because the electric field is zero there), so the probe may be adjusted for high sensitivity for determining that data accurately. However, the probe will affect the shape of the standing waves, with distortion increasing with amplitude, leading to errors in measuring the value (and location) of the peaks. The amount of asymmetry in amplitude-vs.-position provides a qualitative assessment of probe disturbance.

2.1 An example

Having described the method and equipment, it's helpful to go through an actual numerical example to elucidate the procedure.

Step 1: Establish the location of the reference plane: Connect a short-circuit load and note the location of the minima (which should be nearly nulls if the short is reasonably good, and if line losses are negligible). Feel free to increase the probe depth for greater detector output to allow a more accurate pinpointing of the null locations.

Assume for our example that these minima occur at z = –1mm, –121mm, and –241mm. Note also that the wavelength is twice the distance between nulls, or 240mm.

Step 2: Replace the short with the impedance to be measured. Withdraw the probe enough to reduce distortion of the pattern (as evaluated by symmetry), and also verify that the probe output is small enough to lie within its calibrated range. Readjust probe position if necessary to satisfy both requirements. Note both the voltage SWR and the new locations of the minima. If, as is generally the case, the probe produces an output voltage proportional to power, don't forget to compute the SWR by taking the square root of the ratio of the probe output at the maxima and minima.

Assume for our example that the measured SWR is 1.6 and that these new minima are located at z = –41mm, –161mm, and –281mm.

Step 3: Choose one of the null positions from step 1 as the origin, and calculate the difference between this coordinate and the corresponding minimum observed with the load connected.

Here, choose z = –1mm as the origin (it's the closest to the load). Then the displacement we use in the calculation of ϕ is –41mm – (–1mm) = –40mm. Given a wavelength of 240mm, we compute ϕ as

$$\phi = (2n+1)\pi - 2\beta z = \pi - \frac{4\pi}{240\text{mm}}(-40\text{mm}) = \frac{5\pi}{3} \tag{13}$$

where we have arbitrarily chosen n = 0.

Step 4: Compute Γ.

First find $|\Gamma|$ from the SWR measurement:

$$|\Gamma| = \frac{\text{SWR} - 1}{\text{SWR} + 1} = \frac{0.6}{2.6} \approx 0.23 \tag{14}$$

Next, use the phase angle calculated in step 3 to complete the calculation of Γ:

$$\Gamma = |\Gamma| e^{j\phi} \approx 0.23 e^{j5\frac{\pi}{3}} = 0.23\left[\cos\frac{5\pi}{3} + j\sin\frac{5\pi}{3}\right] \approx 0.115 - j0.2 \tag{15}$$

Step 5: Use Eqn. 2 to compute the normalized impedance:

$$Z_{Ln} \approx \frac{1 + (0.115 - j0.2)}{1 - (0.115 - j0.2)} = \frac{(1.115 - j0.2)\ (0.885 - j0.2)}{0.885^2 + 0.2^2} \approx 1.15 - j0.486 \tag{16}$$

Then multiply by Z_0 (here assumed to be 50Ω) to find the load impedance at last:

$$Z_L \approx 57.5 - j24.3 \tag{17}$$

We see that the load impedance (at this frequency) is equivalent to a resistance in series with a moderate capacitance.

And that's all there is to it (more or less).

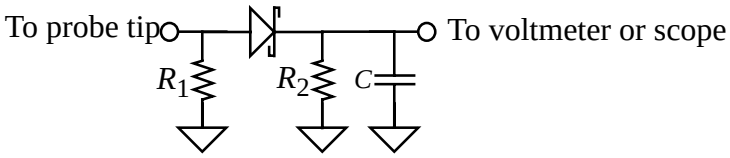
From the foregoing example, it should be clear that the slotted line method involves a fair amount of effort to characterize impedance over a broad frequency range. This is one reason that this method is used less frequently today, although it continues to live on in millimeter wave work where VNAs are either prohibitively expensive or simply unavailable, or where fixturing discontinuities may obscure measurement. It remains without question the best option for hobbyists or labs on a budget, as slotted line gear is readily available on the surplus market at low cost. As an even lower cost alternative, instructions on how to build a simple microstrip-based "slotted" line instrument are given in Section 5.0.

2.2 Error sources (and their mitigation)

Mechanical imperfections are one source of error. For example, if the center and outer conductors are not perfectly cylindrical and truly concentric, the impedance of the line won't be independent of position. Similarly, if the probe carriage assembly does not maintain a constant distance from the center conductor, a position-dependent error will arise. Finally, the dielectric supports that are necessary for mechanical stability inevitably disturb the field patterns as well. In Figure 2 the support is shown as continuous along the bottom, but periodically distributed posts, spaced as far apart as is consistent with providing adequate mechanical support, are also frequently used to minimize perturbations. In any case, the best slotted lines are superb examples of mechanical engineering, with near perfect concentricity. Many are equipped with verniers to allow position measurement with a precision of better than 25μm.

Another (and generally dominant) source of error is associated with the characteristics of the probe. Most probes are simple diode circuits intended to behave approximately as square-law detectors. They thus generate an output voltage roughly proportional to power.

FIGURE 3. Schematic of typical probe



The resistor R_1 is not a physical component of slotted-line probes. It is shown in the schematic simply to remind us that the voltage being sampled by the probe is that of a transmission line, whose impedance is about Z_0 (assuming that mismatches are small).

To gain a crude understanding of the attributes and limitations of a diode as a power detector, assume that the load capacitor in Figure 3 appears as such a low impedance at RF that negligible voltage appears across it. Further assume that the diode continues to exhibit an exponential relationship between current and voltage, even in the RF regime:

$$i_D = I_S\left(exp\left[\frac{qv_D}{kT}\right] - 1\right) = I_S\left[\frac{qv_D}{kT} + \frac{1}{2!}\left(\frac{qv_D}{kT}\right)^2 + \frac{1}{3!}\left(\frac{qv_D}{kT}\right)^3 + \frac{1}{4!}\left(\frac{qv_D}{kT}\right)^4 + ...\right] \quad (18)$$

Next, let the diode voltage (which is equal to the probe voltage with the given assumptions) be a sinusoid:

$$v_D = V_p \sin\omega t \quad (19)$$

The diode current will consist of even and odd harmonics of the input frequency as a result of the nonlinearity. All of these harmonics have a zero time average, so the only contribution to a DC diode current is from the zero-frequency component. Only the even-order terms in the expansion of Eqn. 18 produce DC components, so

$$\langle i_D \rangle = \langle I_S\left[\frac{1}{2!}\left(\frac{qv_D}{kT}\right)^2 + \frac{1}{4!}\left(\frac{qv_D}{kT}\right)^4 + ...\right]\rangle \quad (20)$$

Clearly, the quadratic term is the one that provides an average diode current proportional to the square of the voltage, or proportional to power. All other terms contribute error, with an increasing prominence as the voltage increases. If we are arbitrarily willing to tolerate, say, a contribution from the cuartic term as large as 5% as that from the quadratic, then we must satisfy

$$\frac{1}{4!}\left(\frac{qv_D}{kT}\right)^4 < \frac{1}{20}\left[\frac{1}{2!}\left(\frac{qv_D}{kT}\right)^2\right] \quad (21)$$

or equivalently,

$$\left(\frac{qv_D}{kT}\right)^2 < 0.6. \tag{22}$$

Therefore, as a very crude approximation we must limit peak diode voltage to values

$$v_D < \sqrt{0.6}\frac{kT}{q} \approx 20\text{mV}. \tag{23}$$

Although practical diode detectors vary considerably in their characteristics, it is generally the case that one should probably distrust output voltages readings when they exceed about 5-10mV (perhaps corresponding to input powers on the order of –20dBm or thereabouts). Deviations from ideal behavior increase rapidly as the output voltage increases because the higher-order even terms rapidly increase in significance. The useful range can be extended through the use of a resistive load, at the expense of reduced output level. To understand why the simple trick of resistive loading should be effective, note that the peak *open*-circuit output voltage approaches the input voltage at high amplitudes, behavior which is linear (and therefore clearly sub-quadratic). Loading the circuit with a resistor produces a condition intermediate between the short-circuit case (where the current grows too fast with large input amplitudes) and the open-circuit case (where the current doesn't grow fast enough), leading to a significant range extension. Best results are typically found with a load within a factor of two of 470Ω, with the optimum found by experiment. With the proper load, the acceptable input power range can be extended another 10dB or more. The reduction of output level, however, produces a tradeoff between sensitivity and accuracy.

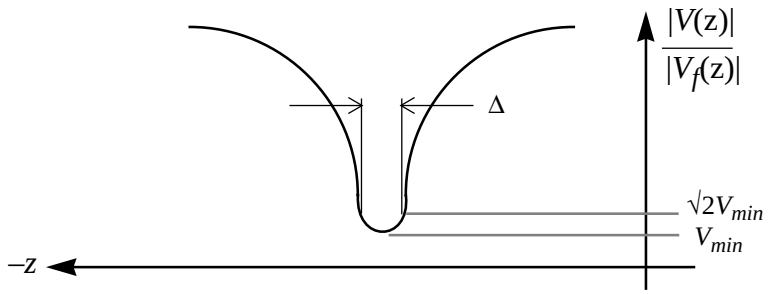
From the foregoing, it is clear that minimizing the peak voltages applied to the detector improves accuracy. However, for a given level of sensitivity, reducing the peak level implies a desire to minimize the voltage ratio to be measured by the detector. If we measure the minimum line voltage, as well as some voltage other than the maximum (as well as the position at which this other voltage is measured), we can accomplish precisely this reduction in the dynamic range required of the detector. This other voltage can be related mathematically to the maximum because the precise shape of the standing wave is known. Specifically it can be shown that the following relationship holds:[3]

$$\text{SWR} = \frac{\lambda}{\pi\Delta} \tag{24}$$

where the quantity Δ is as defined in the following figure:

3. See e.g. Terman and Pettit, *Electronic Measurements*, McGraw-Hill, 2nd ed., 1952, p. 140. This method is described also in *Microwave Measurements*, vol. XX of the MIT Radiation Laboratory Series, McGraw-Hill, 1948.

FIGURE 4. Alternate measurement method for high SWR



This method requires only that the probe accurately measure a voltage ratio of about 1.41:1 (corresponding to a power ratio of 2:1). It is especially useful when attempting to measure very high SWR values, where the maximum-to-minimum voltage ratios are large.

One obvious way to improve accuracy is simply to calibrate the probe to determine explicitly the actual relationship between input and output. However engineers, being the lazy (oops, *efficient*) lot they are, have devised clever workarounds that completely bypass the need to calibrate a probe altogether. Since all that is required is a ratiometric measurement, consider interposing a calibrated attenuator between the signal generator and the slotted line. The attenuation is set to its minimum value (say), the probe is slid along the line until a minimum is found, and the voltage there is noted. The probe is then moved to find the maximum, and the attenuation factor increased until the output is the same as at the minimum. This attenuation factor is precisely the desired ratio V_{max}/V_{min}. Note that this measurement places essentially no demands on the probe at all, having replaced with a readily-realized linear attenuator the need to characterize a nonlinear probe.

Finally, a considerable improvement in sensitivity is possible if the signal generator produces a modulated output. Rather than measuring the DC output of the probe, a demodulator, followed by a bandpass amplifier, provides the output. Using a modulating frequency well above the 1/*f* noise corner of the system improves SNR, allowing the use of higher post-detector gain and a consequent reduction in the required level of coupling of the probe to the line. The reduction in perturbation improves accuracy.

3.0 The Vector Network Analyzer

3.1 Background

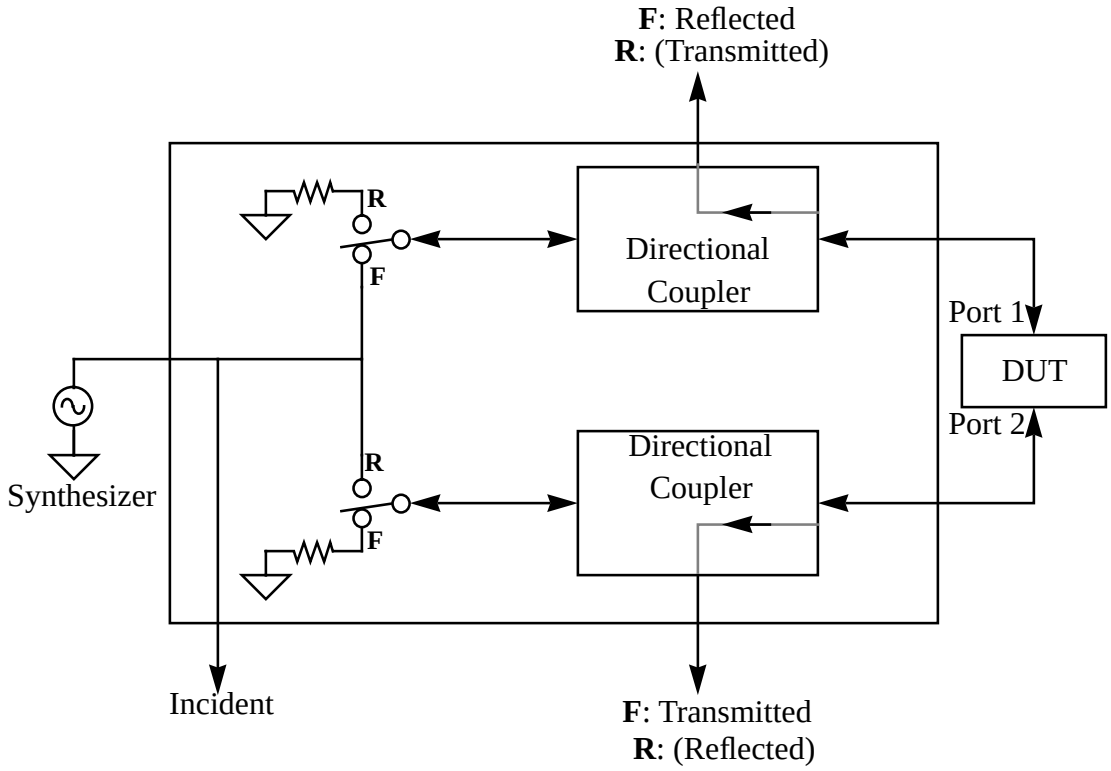
Each data point with a slotted line requires setting the frequency to the desired value, locating the new reference null with a shorted load, and then measuring SWR and locating the minima with the DUT connected as load. The vector network analyzer (VNA) automates this process, and adds greater functionality as well, permitting a rapid and complete characterization of all of the s-parameters of a microwave system over an exceptionally broad frequency range (e.g., from 50MHz to 110GHz in one instrument!).

At the heart of the VNA is a device (e.g., a directional coupler) that miraculously resolves signals along a line into its forward and reflected components. This decomposition into the

two components is valuable because the measurement of impedance can be reduced to measurement of reflection coefficient, as we've seen with the slotted line measurement method. Similarly, measurement of power gain involves ratios of forward components, and so on. Thus a VNA can characterize the full set of s-parameters for a two-port.

A representative block diagram of a VNA reveals the central role of the directional coupler (or equivalent):

FIGURE 5. Typical VNA core block diagram (simplified)



As can be seen in the figure, a frequency synthesizer provides the input to the network analyzer. Both the output power and frequency are controllable. A part of the synthesizer output is sampled as the incident signal, and the rest is steered by a pair of SPDT switches. When the switches are in the position marked "F," the DUT is driven in the forward direction, and the top directional coupler provides an auxiliary output that corresponds to the signal reflected from port 1 of the DUT. At the same time, the lower directional coupler provides an output corresponding to the power coming out of port 2 of the DUT.

To make measurements of reverse characteristics, the switches are moved to position "R," reversing the roles of ports 1 and 2 of the DUT.

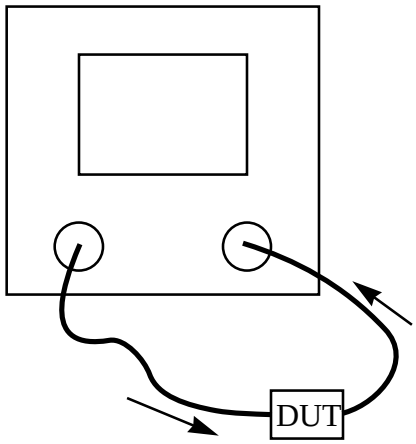
The incident, reflected and transmitted signals are sent to a receiver/detector (not shown) whose job is to measure the magnitude and phase of these signals, followed by processing of the data and presentation in a display.

It is clear that a VNA comprises all of the building blocks of a complete transceiver, and more. Not only does the VNA have to cover an exceptionally wide range of frequencies (e.g., 50MHz to 110GHz in one instrument, albeit with degraded characteristics in the lower decade), but it must make measurements of signals spanning a wide dynamic range of amplitudes at the same time. Operation over such a wide range requires identifying and correcting as many sources of error, both external and internal, as possible. The modern VNA employs sophisticated computational means to accomplish this error reduction, but requires a knowledgeable operator to ensure that the calibration is performed correctly. Mistakes in calibration are an all too common source of anomalous results, so we will spend considerable time examining the error sources associated with the VNA's various measurement modes.

3.2 Basic measurement modes and error sources

First consider making a transmission measurement (either in the forward or reverse direction). There will generally be some cable and fixturing external to the VNA. The total electrical length and loss of these external elements are variable. More to the point, they are beyond the control of the VNA because fixturing is a prerogative of the user.

FIGURE 6. Transmission measurement



One basic calibration step is therefore the measurement of fixturing loss and delay so that these can be subtracted from a subsequent measurement performed with the DUT in place. This step, called the *through* (often abbreviated as "thru") measurement, involves removing the DUT and connecting the rest of the fixturing together directly. The VNA then measures the fixture's phase shift and loss over the user-specified frequency range, storing the data for later subtraction.

After a through calibration, the DUT is inserted and the VNA is ready to measure its insertion loss and phase shift. In many cases, one is interested in the time delay rather than phase. Since delay is simply (minus) the derivative of phase with respect to frequency, the VNA can readily compute the delay from phase data. There are some subtleties, however, that one must appreciate if correct measurements are to be made. One such consideration is that the instrument measures phase at a discrete set of frequencies, rather than continu-

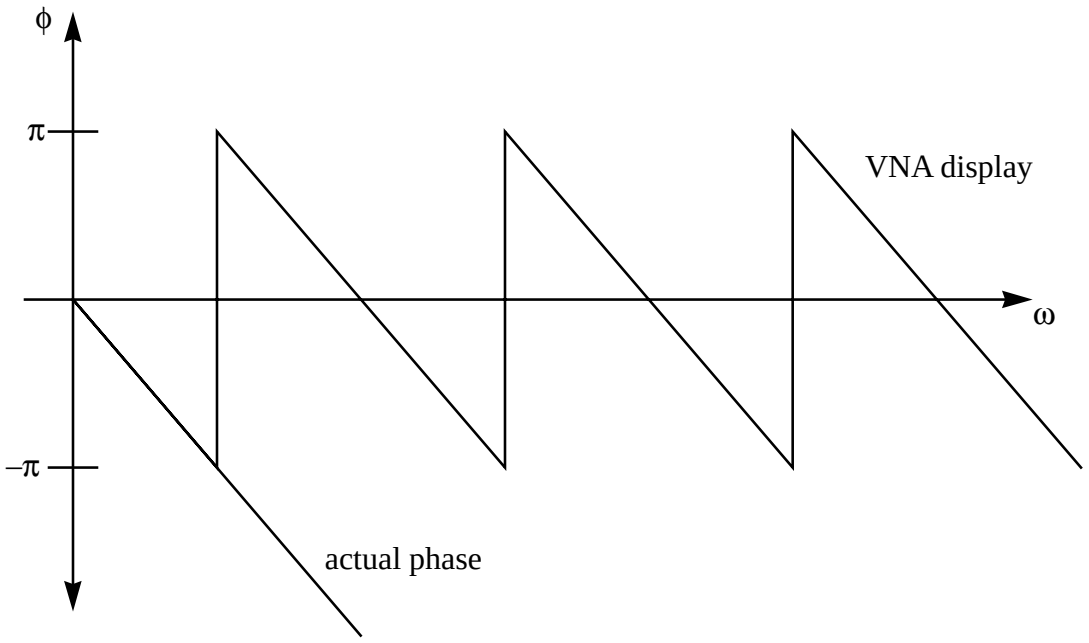
ously over the entire band. Hence, the derivative must be approximated by a ratio of finite differences:

$$\frac{d\phi}{d\omega} \approx \frac{\Delta\phi}{\Delta\omega} \tag{25}$$

The frequency interval in the denominator of Eqn. 25 is known as the frequency *aperture*, and is controllable by the user. A narrow aperture provides fine resolution, but may be sensitive to noise in the data. A wide aperture is less sensitive to noise because it effectively performs an averaging over the frequency interval, but can miss fine structure precisely because of this averaging. Modern instruments default to an aperture that is satisfactory for most applications, but which may be overridden by the user if desired.

Another subtlety is that the phase detector within a VNA functions over a finite interval, modulo some phase. A typical detector range is $\pm\pi$ radians, so the VNA cannot distinguish phase shifts outside of this range from those lying within it. Hence, a pure time delay's phase appears as a periodic sawtooth when plotted against frequency in linear coordinates:

FIGURE 7. Phase shift vs. frequency, ideal vs. VNA display



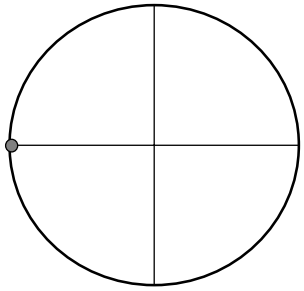
The user must employ physical arguments or other knowledge to splice the various regions together properly; the VNA fundamentally lacks the information necessary to do so. There is an advantage to the sawtooth-like display however, when plotting, as it reduces the total vertical height for a given resolution.

A related consequence of the modulo-ϕ behavior is that if a VNA's computation of delay uses an aperture value that corresponds to a phase step in excess of π radians, the displayed delay will be in error. To guard against these types of problems, it is good practice to examine both the phase and delay curves, rather than just the delay. Simple checks of

reasonableness are carried out rapidly, so there is hardly an excuse for not performing them.

In addition to transmission measurements, the other basic VNA operating mode is measurement of reflection. Just as with the slotted line, it is necessary to establish a reference plane. And, just as with the slotted line, a short-circuit load can be used for this purpose. So, in the simplest calibration for a reflection measurement, the best available short-circuit load is connected to the test port in place of the DUT. The VNA measures the magnitude and phase of the reflection over the specified frequency range and stores this data, using it to locate the reference plane and correct for fixturing losses (the VNA cannot use the information about fixturing losses from the through measurement because the latter does not identify the loss over the relevant fixturing path length). After this calibration step, a display of S_{11} with the short-circuit load should consist of data points tightly clustered about the –1 point:

FIGURE 8. VNA display after calibration with short



If other than a tight distribution is observed (e.g., an arc), carefully check the fixturing (particularly the connectors), correct any problems, and repeat the calibration. After re-verification, the VNA is ready to perform one-port reflectance measurements. As we saw with the slotted line, such a measurement is equivalent to an impedance measurement. Depending on context, the user may wish the data to be displayed as reflection coefficient or impedance. The modern VNA can provide both a display of Γ in polar form, or impedance on a Smith chart. The format is deliberately left unspecified in Figure 8 because for the special case of a shorted load, the data are located in the same spot.

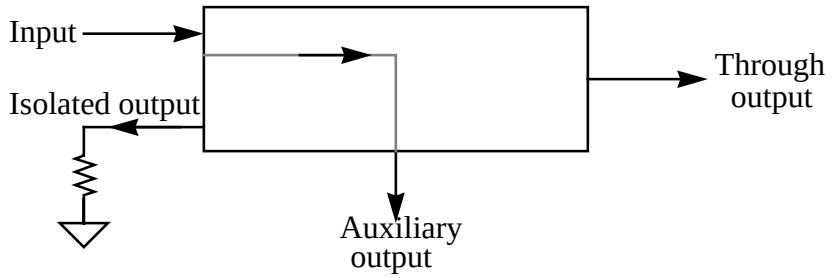
A subtle issue is that the calibration with the shorted load establishes the reference plane at the physical location of the short within the calibration standard. The fixturing may add some physical length beyond that plane. One could correct for this by repeating the calibration with a short at the actual DUT terminals. Another option is to make use of the "port extension" feature of modern VNAs in which the instrument algorithmically adds length, effectively moving the reference plane further away from the VNA connectors. The correct extension is determined by producing the best possible short at the DUT terminals, and varying the extension value to minimize the size of the distribution near the –1 point on the Smith chart.

The foregoing description focuses on how external fixturing errors can be removed. Using the through and short calibrations, the VNA can reduce by large amounts the errors in transmission and reflection measurements. For even greater accuracy, the VNA is capable

of characterizing its own internal errors with the aid of additional calibration steps. To appreciate how the VNA performs these additional corrections, it is necessary to identify the errors corrected in these various calibrations.

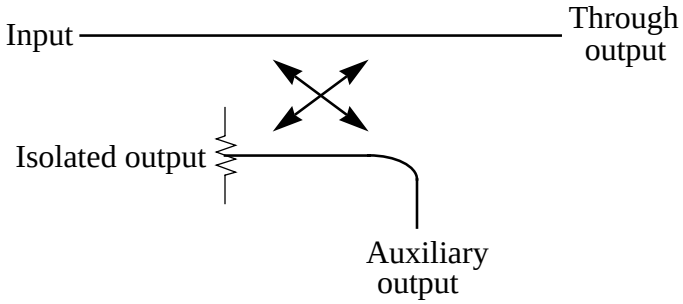
The VNA depends on directional couplers to decompose signals into incident and reflected components. As with everything else, practical directional couplers are imperfect. To quantify these imperfections, we need to define the various figures of merit which apply to the directional coupler as it is configured for use in a VNA:

FIGURE 9. Directional coupler port definitions



It should be noted that Figure 5 and Figure 9 use a simplified symbol for the directional coupler. One more commonly used symbol is

FIGURE 10. A more standard symbol for a directional coupler



From both symbols, it is clear that the directional coupler is generally a 4-port device, but a VNA typically uses only three of them, terminating the fourth (the isolated output) in a matched load.

Initially assume that the through output is terminated properly. Most of the power supplied to the input port travels on to the through output, with a small portion coupled to the auxiliary output. One characteristic parameter is therefore the coupling factor, defined as the ratio of input power to auxiliary power:

$$C \equiv \left. \frac{P_{IN}}{P_{AUX}} \right|_{Forward} \tag{26}$$

Typical values of coupling factor range from 3 dB up to a bit over 20dB. A larger coupling factor means that more power is coupled to the main (through) output, not to the auxiliary output. Therefore the lower the coupling factor, the greater the loss in going from the input to the through output.

If the directional coupler is operated in reverse, with power now supplied to the through output with the input terminated, ideally no signal should be measured at the auxiliary output (that's the reason for the "directional" nomenclature). Inevitably, though, some reverse power will leak through to the auxiliary output. A measure of how well the reverse leakage is suppressed is the isolation factor, defined as

$$I \equiv \left. \frac{P_{IN}}{P_{AUX}} \right|_{Reverse} \tag{27}$$

Isolation factors of 30-60dB are not uncommon.

The two quantities are often combined to yield a figure of merit called the directivity, *D*:

$$D \equiv \frac{I}{C} = \frac{\left. P_{AUX} \right|_{Forward}}{\left. P_{AUX} \right|_{Reverse}} \tag{28}$$

Thus directivity is a measure of how well the coupler discriminates between forward and reverse components. We desire an infinite directivity, but all real couplers fall short of the ideal. In practice, directivities of 20-40dB are typical. The lack of infinite directivity is a significant error source, and correcting for this deficiency is a major aim of VNA calibration.

To illustrate how directivity errors can corrupt measurements, first examine Figure 5 and Figure 9 to review how directional couplers are hooked up inside a VNA. Notice that the main input of each directional coupler is connected to a port of the DUT. Thus when performing a reflectance measurement the synthesizer drives the through output. Power flows from the synthesizer, "backwards" through the coupler, to the DUT. Any power reflected by the DUT feeds back into the main input of the directional coupler, and a portion of the reflected power exits the auxiliary port for sampling and measurement.

If the directivity were infinite, the auxiliary port signal would be due entirely to the power reflected from the DUT, allowing direct measurement of the reflected power. However, a finite directivity implies that some of the power flowing from the synthesizer to the main input leaks out of the auxiliary port as well. The VNA would then measure an auxiliary port signal that is a weighted sum of both the forward and reflected power. Because these may add both in and out of phase over frequency, typical manifestations of imperfect directivity are ripples in, say, the measured reflection coefficient as a function of frequency.

A representative calculation illustrates the magnitude of the problem. Suppose we have a 10dB coupler with 30dB directivity, and we attempt to measure the impedance of a load that has a 20dB return loss. That is, C = 10dB, D = 30dB, $I = C + D$ = 40dB and RL = 20dB. The signal reflected by the DUT is RL = 20dB below the incident power level, and the amount of the reflected signal surviving to the auxiliary output is C = 10dB below that, for a total of 30dB below incident. The unwanted signal at the auxiliary port is I = 40dB below incident. Thus we see that the error power is an unacceptable 10% of the signal power in this example. If we were to attempt to measure a return loss of 30dB, the situation would be even worse, for the error power would then equal the desired signal power.

Another error that behaves much like directivity error arises from reflections at interfaces with adapters, cables and fixturing. These reflections necessarily produce signals at the auxiliary output of the coupler and, as in the previous example, these parasitic signals can obscure the component of signal that is due to the actual reflection from the DUT.

Source mismatch is yet another potential source of error. Consider the flow of power from the source, through the coupler, and to the load. Some power reflects off of the load and returns to the source. If there is a mismatch in source impedance, there will be a subsequent re-reflection from the source back through the coupler. Some of that power reflects off the DUT, and finds its way out of the auxiliary port. From the qualitative description of this process, this error term is clearly most significant when the load has a high reflection coefficient.

A third type of error is related to one we've already examined: frequency response. The couplers, cables, adapters and the part of the system that actually measures magnitude and phase may all have frequency dependent characteristics.

The three types of one-port errors – directivity, source mismatch and frequency response – can be removed by performing three experiments. For example, consider attaching a perfectly resistive matched load as the DUT. In this case, the auxiliary output of the directional coupler should have no signal. Any deviation from that condition indicates an effective directivity error, which can be measured and stored for later removal. The extent to which directivity errors are nulled out depends critically on the quality of the "perfect" load used in this step of the calibration sequence.

A common choice for the other two experiments is to use both a shorted and open load. As with the perfect load, the ultimate accuracy of VNA measurements depends on how close the impedance of the loads are to zero and infinity. It is particularly hard to implement a good open-circuit at high frequencies because stray capacitance is difficult to control. To underscore the difficulty involved, note that a 0.1pF stray capacitance (which is about the value of an open-circuited APC-7 connector) has an impedance of only about 160Ω at 10GHz. Also, radiation from the open end is also an increasing problem as frequency increases, and this loss produces a real component of impedance in parallel with the parasitic shunt capacitance. This problem is mitigated by sliding a short along a line until it is positioned a quarter wavelength away from the reference plane. The shorted line is a closed structure which prevents radiation.

When calibrating for two-port operation, the three experiments are augmented with a fourth: a through measurement to characterize the frequency response of the fixturing, as described earlier. Hence, the quartet of experiments is often known as the “short/open/load/thru” (SOLT) two-port calibration method.[4]

There are several minor variations on the SOLT technique, all aimed at solving the problem of imperfect impedance standards. One of these replaces the fixed matched load with a sliding load. Here, a movable load with a near-perfect match is slid along an air line, and the whole assembly used in place of the fixed load. As the load slides along the line, the small reflections combine with the incident signals in a periodic manner, alternately adding and subtracting, leading to a data set that is distributed in a circle in the complex plane. The directivity vector is the center of that circle. Three points uniquely determine a circle, so in principle only three measurements are needed. In practice, a larger number is used to improve the error estimation. The sliding load is generally considered necessary above approximately 2-3 GHz.

Yet another variation replaces the sliding load with an offset load, which may be thought of as a sliding load in which the load no longer slides. If two points and the angle of the offset are known, the center of the circle can again be determined. These two points are obtained with two loads of different length. The sliding or offset loads are popular at millimeter wave frequencies where ideal loads are simply not available. At these frequencies, a shim of known thickness is inserted between mounting flanges to produce the second measurement.

An alternative calibration suite is known as the thru-reflect-line (TRL) method. This method corrects for the same errors as the SOLT method, but depends less on the perfection of the impedance standards used as calibration loads. As the name of the method suggests, the first step in the calibration is to connect the two ports of the external fixture together in a low reflectivity through configuration to characterize the fixturing. The next step is to connect a grossly mismatched load to each port of the fixture separately (hence the name “reflect”). The precise nature of the mismatch is irrelevant, and its magnitude need not be known. It just has to have the same high reflectance at both ports (a nominal short is frequently used). Finally, the two ports are again connected together through the low-reflectivity fixturing, but now with a different cable (or other fixturing) length than was used in the through measurement.

The TRL method is particularly attractive for non-coaxial systems such as microstrip, where impedance standards are difficult to realize (or are simply unavailable commercially). It is also attractive for coaxial media because impedance standards are expensive, and the TRL procedure requires no expensive elements. However, the TRL calibration only works over an octave range for a given length of cable, owing to the periodicity of reflections. Progressively longer fixturing is needed to extend the TRL calibration to lower and lower frequencies.

4. This is also known as SLOT, LOST, SOT-L and a variety of other permutations.

4.0 Special Considerations for Microstrip

As implied in the preceding paragraph, microstrip environments pose some challenges for calibration. Consider, for example, a microstrip fixture for measuring the s-parameters of a transistor. It is not entirely obvious how to carry out, say, a SOLT calibration sequence for such a non-coaxial structure. Since one important aim of calibration is to null out fixturing artifacts, we evidently wish to implement and use a short, open, matched load, and a through line at various stages of the calibration, all at the physical location where the DUT (here, a transistor) would be placed. An open circuit sounds easy enough, and so does a through. The former can be approximated simply by not installing a DUT, and the latter can be approximated by a second fixture identical to the first, but in which the microstrip line extends all the way across. Implementing a reasonable approximation to a short is similarly straightforward, with a third fixture (again otherwise identical to the first) in which the ports are shorted to ground (e.g., through a nice line of vias). The tough one is implementing a good matched termination. A surface mount resistor at the end of the microstrip line, for example, might suffice for approximate work, for is unsatisfactory for accurate characterization. Its series inductance and shunt capacitance cause the impedance of the “matched” load to vary over frequency.

A reasonable solution to this problem becomes apparent when we re-examine what errors are being calibrated out with the matched load connected to the VNA. For the most part, internal VNA directivity errors are being nulled out at this step of a SOLT calibration, so there is no need for the rest of the fixturing to be involved at all. Hence, the ordinary coaxial matched impedance standard may be connected directly to the VNA port without worrying at all about whether a microstrip test environment will eventually be used. We will call this method the modified SOLT technique.

The modified SOLT calibration unfortunately will not fix errors in effective directivity caused by a mismatch past the APC port. Hence, if that transition is poor, or if precise answers are necessary, then a TRL calibration should be performed. For exacting work, then, the TRL calibration method is better, but the modified SOLT is often good enough.

A final consideration is that the “open” condition with a microstrip is imperfect (as it is with all open structures) because of fringing capacitance (on the order of 50fF). A partial correction for this is possible through the use of software connector subtraction. Many VNAs have the ability to remove the effect of a connector through software means. Alas, microstrip is not one of the ordinary options. Of the options that are typically available, the best approximation is the APC, whose ~100fF fringing capacitance is reasonably close to that of a typical open line on FR4. A residual error remains, however, and one may use the variable port extension feature of many VNAs to reduce this residual error substantially.

5.0 Summary of Calibration Methods

The foregoing section describes so many permutations that it is easy to get a bit confused (and to make things even more confusing, we haven’t covered all of the ones that exist).

Here is a summary of the calibration methods, along with some comments to remind you what their relative attributes and weaknesses are, allowing you to make an informed choice of calibration technique.

The simplest is the short-thru one- and two-port calibration, which only corrects for external fixturing and detector frequency response errors. Errors from finite directivity and source mismatch are not corrected. This technique is also known as *response calibration*.

Better for one-port measurements is the short-open-load (SOL) suite of calibrations. As long as the impedance standards are perfect, this method is capable of nulling out errors from finite directivity, as well as source mismatch and detector frequency response errors.

A thru measurement may be added to yield a SOLT calibration which is a two-port method that corrects for all of the errors corrected by SOL, and also corrects for the remaining cabling of a two-port fixture.

Variations on the basic SOLT theme include replacing the load measurement with either a sliding load or a fixed offset load. The modified SOLT method employs an SOT sequence with a microstrip (or other non-coaxial) fixture to null out all but VNA directivity errors. Use of a standard coaxial matched load without the fixture completes the calibration by (almost) zeroing out directivity errors.

Finally, the thru-reflect-line (TRL) method eliminates the need for perfect impedance standards as calibration loads while correcting the same errors as the SOLT technique. The TRL method is particularly attractive in characterizing non-coaxial systems such as microstrip, but only functions over an octave of frequency for a given fixturing length.

With TRL calibration, it is possible to reduce directivity and source mismatch errors to levels as low as –60dB at 18GHz, and essentially eliminate frequency response errors. These values should be compared to –40dB directivity and –35dB source mismatch errors typically achieved with the SOLT method (fixed load). The sliding and offset load options improve the SOLT errors to levels in between those of fixed SOLT and TRL methods.

6.0 Other Measurements

Thanks to the extensive use of computation, the modern VNA is capable of more than a complete measurement of s-parameters. For example, once the s-parameters are measured over a broad frequency range, the frequency response data can be transformed to time response data. Step responses and TDR traces can be generated from VNA data. Although the time taken to perform all of the measurements and computations is substantially larger than it would take for a “real” TDR, this additional functionality is nonetheless welcome.

Because of the algorithmic nature of the transform, it is possible to perform a little mathematical magic that would be impractical to carry out with actual time-domain instrumentation. For example, consider a case where a TDR trace contains reflections from multiple sources. The early discontinuities can mask the effect of subsequent ones in a real TDR measurement. A VNA, however, can remove the first discontinuity, allowing examination

of the previously masked reflections. The extent to which a VNA can perform this removal (called *gated impedance* or *gated TDR* measurements) depends on the accuracy and noise of the s-parameter measurements.

7.0 References

Aside from the sources cited in footnotes, the reader may also find useful *Vector Measurement of High Frequency Networks* (Hewlett-Packard High Frequency Vector Measurement Seminar notes, April 1989). These notes contain an excellent high-level summary of how a VNA is used, with a concise discussion of error sources and calibration methods. Another useful reference is the user manual of almost any VNA, such as the HP8720C (a 130MHz to 20GHz instrument) or the HP8510C (capable of operating from 50MHz to over 100GHz).

8.0 Appendix: Microstrip "Slotted" Line Project

As we've seen, the modern network analyzer is a truly remarkable instrument, capable of extremely accurate characterizations of microwave networks over a broad frequency range. Unfortunately, this capability comes at a price: A typical GHz VNA costs more than the average sports utility vehicle, and the few that are available on the surplus market are rarely significantly discounted. Clearly, a VNA is generally priced out of the reach of most hobbyists (and even out of the reach of many academic laboratories), so the slotted line is the device of choice for those on a budget. On top of that, the slotted line is a superb pedagogical tool for teaching the principles of Smith chart manipulations (e.g., providing explicit explanations for phrases such as "wavelengths toward the generator" and so forth). As mentioned earlier, many slotted lines are available on the surplus market for quite reasonable prices, at least for lines designed for use in the low GHz frequency range.

This section describes a much cheaper, and much cheesier alternative: a microstrip "slotted" line system capable of functioning to 5GHz and beyond. This instrument (and that is a loose use of the term, to be sure) has important attributes: it costs very little (the total parts and materials cost should not exceed $5-$10) and is extremely easy to make using ordinary tools and materials. The tradeoff is that the instrument's accuracy is not particularly good, and the shaky mechanicals are not terribly robust. However the performance is adequate for virtually any home project in the low-GHz range of frequencies.

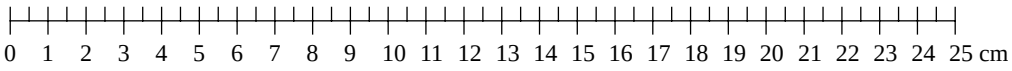
The design presented here is based on FR4 material and right-angle mounted bulkhead BNCs of the type described in the chapter on microstrip, in keeping with a focus on minimizing cost. In particular, a microstrip line is much easier to make than a slotted coaxial airline. Of course, better performance can be obtained by using lower loss PC board material in tandem with better connectors, and the reader is certainly invited to improvise variations on the basic design as budget, patience and performance requirements dictate.

The first step is to get a piece of FR4 longer than the largest electrical wavelength of interest, but not so long that the loss is excessive over the desired operating frequency range.

For a minimum operating frequency of 1GHz, a good compromise is about 25cm. A line of this length typically exhibits about 0.8dB of loss at 1GHz, and perhaps 4dB of loss at 5GHz (compared with under 1dB for a true coaxial slotted line). If you are going to use the instrument only at the higher frequencies, performance will improve by shortening the line to reduce the loss (we only need the line to be long enough to contain a couple of minima, and the loss *per wavelength* is roughly constant, at a value of a bit under 1dB per λ).

Mount BNCs at the two ends, and then construct a 50Ω microstrip line using copper foil tape. It is important that the foil be as smooth as possible. Next, affix a nonconductive metric ruler just below the line (if you don't have a suitable ruler, use a photocopier to duplicate the following metric ruler at twice scale):

FIGURE 11. Metric ruler for microstrip line (drawn at half size)



Because photocopier accuracy varies considerably, verify that the enlargement hasn't distorted the scale factor of the ruler. Careful interpolation between the 5mm markings should allow a precision of ~1mm; accuracy is a different matter!

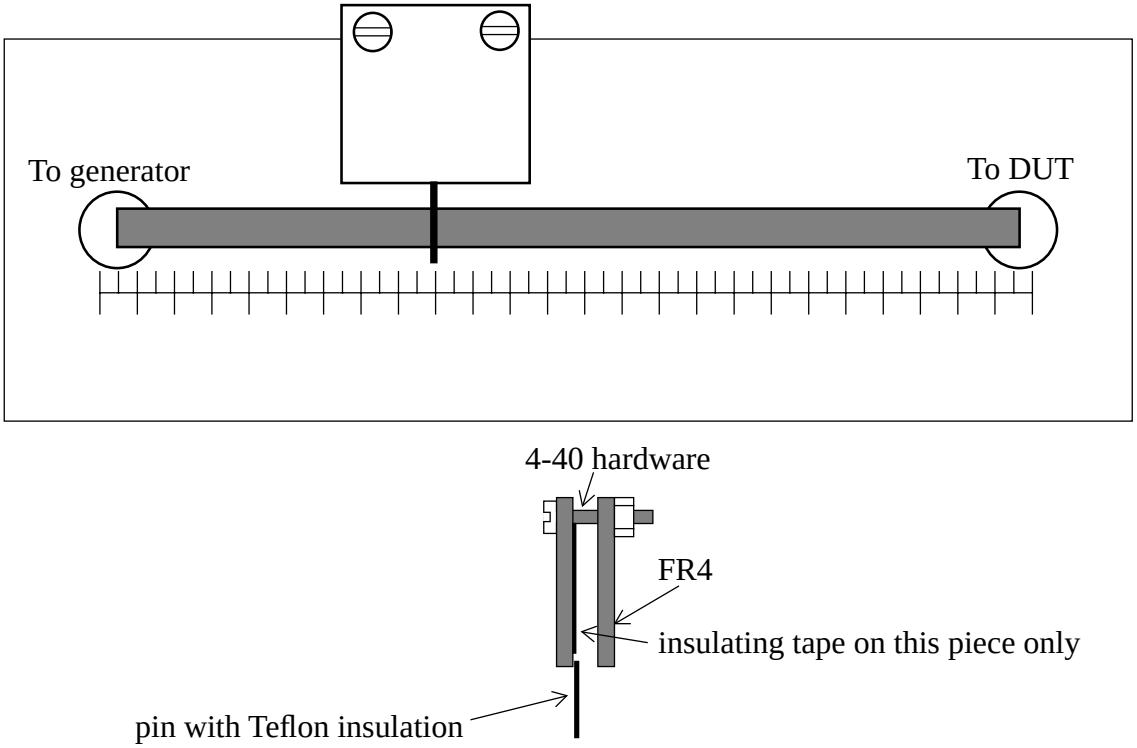
The next step is to construct the detector. Here we use a Schottky diode-based detector circuit capacitively coupled to the microstrip line. It's a simple circuit, and the biggest challenge you'll face is mechanical, to construct the slider while guaranteeing proper and consistent coupling of the detector to the line.

The probe is a common needle, such as the kind that comes with new clothes, carefully jammed into the bottom side of one part of the slider and cemented with a little epoxy, then clipped to length (see Figure 10). The probe is surrounded by a short length of insulation (preferably Teflon) taken from a piece of hookup wire to act as the dielectric between the probe and the line, and also to provide a smooth rolling action. The probe motion (and the line itself) must be as smooth as possible to maintain a constant coupling as the probe slides along the line.

The slider assembly is made out of two pieces of FR4 that are bolted together. Teflon tape (or very smooth copper foil tape) may be affixed to the inner surface of the piece that's on the line side of the unit to reduce sliding friction and abrasion. Its thickness needs to be carefully controlled to ensure that the probe makes good contact with the line.

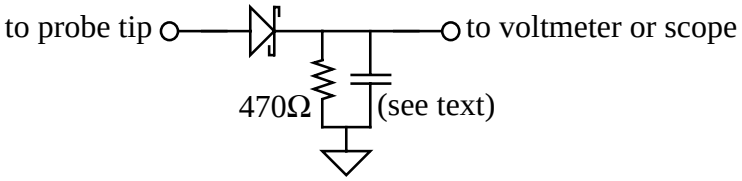
The foil side of both pieces faces the main board. The foil that contacts the ground plane of the main board provides the ground contact for the probe circuitry.

FIGURE 12. Bottom view of slotted line and side view of probe assembly (not to scale)



The slider contains the probe circuitry, which consists of a diode detector and a resistive load:

FIGURE 13. Schematic of probe assembly



The Schottky can be any low-capacitance, high-frequency unit (such as the HP 5082-2835 or -2860), the anode lead of which is connected to the actual probe tip. The input signal amplitudes must be small enough that the diode acts approximately as a square-law detector, making the output voltage roughly proportional to power. The proportionality constant, as long as it truly is a constant, is fortunately irrelevant since only ratios are used in computing SWR. However, the square-law behavior necessitates taking the square-root of the probe output voltages in order to compute SWR.

The resistive load is shown as 470Ω, but you are encouraged to experiment with its value to maximize the detector's useful range. A surface mount chip resistor is used to keep parasitics small.

The capacitance of the slider, plus the input capacitance of most meters and scopes to which the output of the probe is connected, will generally be large enough not to require any additional capacitance. To maximize usefulness, the ability to measure sub-mV signals is desirable. Some amplification may be necessary to boost detector outputs to levels that are conveniently measurable with inexpensive instruments.

So, when all is said and done, how good is the microstrip slotted line? The lossiness of FR4 and the right-angle mounted BNCs, probe coupling irregularity, lack of probe calibration, and the hand-built nature of the line itself all conspire to make this impedance measurement tool a rather crude one. As a rough rule of thumb, one can expect reasonable accuracy for impedances between about $Z_0/5$ and $5Z_0$. For the most common case, that of producing a good match to Z_0, the tool works extremely well, allowing the attainment of S_{11} values below –15dB with ease. Results at 1GHz are generally surprisingly good, with progressive degradation as the frequency increases to 5GHz and beyond.

If the instrument is to be used only at the higher frequencies, there are several necessary refinements. Replace the right-angle BNCs with inline SMA connectors, use RO4003 instead of FR4, and abandon the idea of handcrafting the line out of copper foil tape; it is insufficiently uniform, so conventional PC board manufacturing techniques are required. Finally, a better diode may have to be used (e.g., the M/A-COM MA4E2054, which is specified beyond 10GHz). If the line is shortened by a factor of 5 or so, satisfactory operation between 5 and 10GHz is possible when all of these refinements are combined.

Derivation of Fringing Correction (Danger, Will Robinson – Integrals, Cheese and a Breeze Ahead!)

1.0 Introduction

A rigorous calculation of fringing capacitance is rather difficult. Although there are many clever analytical methods, such as those based on Schwarz-Christoffel conformal mapping techniques, there are often numerous practical restrictions on their applicability. When precise answers are needed, or if the geometry is complex, often the best choice is to employ numerical methods. Unfortunately such an approach often obscures design insight. As a complement to those valuable approaches, we offer an analytical expression whose inaccuracy perhaps can be forgiven in view of its simplicity and near universality. And although its derivation may not exactly fit on a cocktail napkin, the final result certainly does, as will be clear shortly.

The approach we'll take is inspired by one of the many wonderful chapters in Feynman's *Lectures on Physics*, in particular, The Principle of Least Action.[1] There, Feynman points out that elegant and powerful minimum principles can frame novel solutions to old problems. For example, if you were to forget the current divider law for two parallel resistors, you could derive it using the principle that currents will distribute themselves in a way that minimizes the total power dissipation. Any other current distribution would result in a higher total dissipation (try it!).

Similarly, if our task is to deduce the electric field between two conductors, it is valuable to know that charges will distribute themselves to minimize the total energy stored in the system, and also to remember that a unique potential distribution is linked to the charge distribution. Since, for a given voltage, energy is proportional to capacitance we may infer from this minimum principle that the correct potential distribution is the one among all possible distributions that minimizes the computed capacitance.[2] We use this observation by proposing a "reasonable" functional form for the potential distribution, computing the capacitance it implies, and then choosing parameters (if any) to minimize that capacitance. Feynman's minimum principle then says that we will have generated the best possible approximation to the truth, for *that* particular guess (even if it is wrong). Furthermore, we will know that our approximation error will always be positive (our approximate formula will necessarily overestimate the true capacitance), so at least we'll know the sign of the error.

To start, we equate two different formulas for the energy stored in a capacitor:

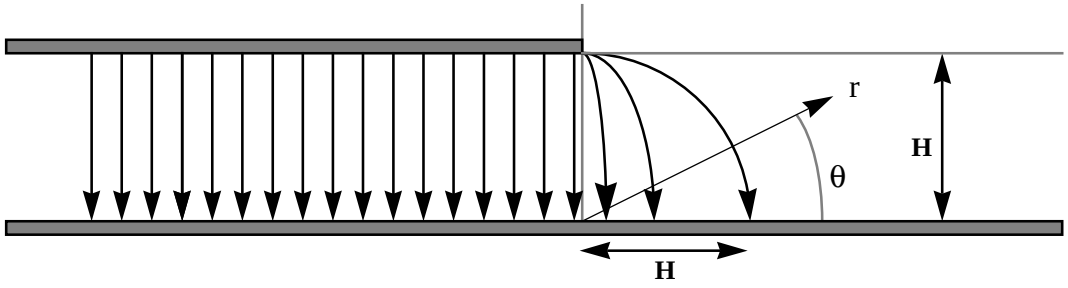
1. Chapter 19, Volume II, (Addison-Wesley, 1964).
2. The same minimum principle can be used to derive formulas for inductance.

$$\frac{1}{2}CV_0^2 = \frac{1}{2}\varepsilon \int_{Vol} |\nabla V|^2 dVol \ , \tag{1}$$

where ∇V is the gradient of the potential V (recall that the electric field is equal to minus this gradient). The term on the left comes from ordinary circuit theory, that on the right from field theory.

Next (and this is the tricky part), guess a "reasonable" form for the potential. To aid in guessing, first look at our structure:

FIGURE 1. Very approximate field distribution for fringing capacitance estimation (side view)



The electric field lines are idealized as perfectly vertical until the very end of the line is reached, then progressively curving outward more and more until they are perfectly circular at a distance H beyond the end. Along the radial line shown in the figure, and at an angle θ with respect to the ground plane, assume that the potential increases in some fashion as the radius r increases from 0 to H. Further assume fancifully that, at a given r, the potential increases linearly from 0 as the angle θ varies from 0 to $\pi/2$. Assume also that negligible energy is stored in the electric field for $r > H$. This latter assumption avoids an embarrassing prediction of potentials in excess of the applied voltage, V_0, as the radius goes to infinity. It also causes us to underestimate the energy stored. This error is at least in the right direction to offset the systematic overestimation inherent in the method when used with any incorrect potential distribution (although there is the possibility of overcompensation). Finally, assume that the plates are infinitesimally thin.

Given these assumptions (and they are just that), we may postulate an approximate potential function of the following form:

$$\tilde{V}(r,\theta) = V_0\left(\frac{r}{H}\right)^k\left(\frac{2\theta}{\pi}\right), \tag{2}$$

where k is some parameter whose value is to be determined later. The tilde denotes that it is a postulated, and approximate, potential. You can verify that this equation satisfies the conditions stated above (but not necessarily all relevant boundary conditions; if it did, it would have to be the correct solution). Note that we neglect variations in the z- direction (out of the plane of the page).

With that potential function in hand, the rest is just plugging and chugging:

$$\tilde{C} = \frac{4\varepsilon}{\pi^2 H^{2k}} \int_{Vol} |\nabla \tilde{V}|^2 dVol = \frac{4\varepsilon W}{\pi^2 H^{2k}} \int_{Area} |\nabla \tilde{V}|^2 r dr d\theta \,, \tag{3}$$

and

$$\nabla \tilde{V} = \theta k r^{k-1} \hat{r} + \frac{1}{r} r^k \tilde{\theta} \,. \tag{4}$$

After combining these equations and evaluating the double integral as r ranges from 0 to H, and as θ varies from 0 to $\pi/2$, we get

$$\frac{\tilde{C}}{W} = \frac{\pi\varepsilon k}{12} + \frac{\varepsilon}{\pi k} \,, \tag{5}$$

where W is the width of the line.

Now, we want to select k to minimize the estimated capacitance. Setting the first derivative of Eqn. 5 to zero yields

$$k = \frac{\sqrt{12}}{\pi}, \tag{6}$$

which, given the approximate nature of this entire endeavor, may be treated as essentially unity (meaning that we could have just started with $k = 1$ and ended up with pretty much the same answer).

Substitution of this value of k into Eqn. 5 finally gives us an approximate equation for the per-width fringing capacitance:

$$\frac{\tilde{C}}{W} = \frac{\varepsilon}{\sqrt{3}}. \tag{7}$$

Note that this capacitance is **independent** of H. To the extent that the approximations leading to its derivation are valid, it is therefore a *universal* fringing correction, whose value is very roughly 5fF/mm (assuming a vacuum dielectric; multiply this figure by the relative dielectric constant in general). That is, any open structure will contribute approximately this capacitance per length of edge, at least to cocktail napkin accuracy.[3]

Since capacitance is proportional to the ratio of effective area WL_{eff} to plate spacing H, the fringing capacitance is equivalent to an ideal fringing-free parallel plate capacitor whose dimensions are W by $H/\sqrt{3}$.

3. This statement is true for a finite-length conductor over an infinite ground plane. For two equal-size plates, the universal correction is precisely half the value, or about 2.5fF/mm. But the length extension remains $H/\sqrt{3}$ (or, after rounding, $H/2$) per edge. Also remember that we have ignored field variations in the z-direction, so the correction gets increasingly dubious as W/H diminishes.

But wait, you say: Our correction has $H/2$, not $H/\sqrt{3}$, in it. Here's how we get $H/2$: First, we know that our proposed functional form is wrong (consider its behavior at large radii). So, by the minimum principle we know that our estimate is probably too high ($\sqrt{3}$ is too low), if the field for $r > H$ were truly negligible. But by how much? We don't know (if we did, we could remove the error altogether). But under the assumption that the estimate isn't too horribly wrong, we arbitrarily round the denominator upward a little bit to get to the closest convenient number, 2. Cheesy? You bet. Can we evaluate the amount of cheesiness? Consider the following table, which presents correction factors for the capacitance between circular parallel plates of diameter D and spacing H. We define a correction factor as the ratio of actual (or estimated) capacitance, to the value given by the simple fringing-free undergraduate physics formula. Values in the second column are obtained from numerical field solutions, and the third from assuming that the capacitors act as if we extended the radius by an amount $H/2$ (so that the effective diameter is $D + H$).

TABLE 1. Circular parallel-plate capacitance

H*/*D	**"Exact" correction factor**	**Cheesy correction factor (using *H*/2)**	**Residual cheese error (%)**	**Cheesy correction factor (using H/√3)**	**Residual cheese error (%)**
0.005	1.023	1.010	1.2	1.012	1.1
0.01	1.042	1.020	2.1	1.023	1.6
0.025	1.094	1.051	4.0	1.059	3.2
0.05	1.167	1.102	5.5	1.119	4.1
0.10	1.286	1.210	5.9	1.244	3.3

As expected the correction factors are very close to unity for small spacings, so all three formulas yield answers that differ negligibly from each other. As H/D ratios grow, however, the fringing-free parallel plate formula underestimates the true capacitance by increasing amounts. The true capacitance is nearly 30% larger than the value computed by the fringing-free formula when the H/D ratio is 0.1. Application of the cheesy correction factor results in a residual error that is under 6% at that same spacing. This close tracking is encouraging, because the correction factor was derived for a rectangular structure, but applied successfully to a circular one. The assertion of a universal fringing correction thus seems less unreasonable.

As a final comment, if we use the $H/\sqrt{3}$ factor actually derived, instead of the $H/2$ arbitrarily substituted for it, the error improves a little bit in the particular case of Table 1, as can be seen in the last two columns.[4] At a normalized spacing of 0.1, the correction factor becomes 1.24, reducing the error to almost a full order of magnitude below the fringing-free estimate, to a little bit above 3%. So, which to use? Fortunately, the contribution by

4. A value of $2H/3$ is even better for this particular data set, with almost no error at $H/D = 0.1$.

fringing itself typically constitutes a second-order correction to first-order formulas, so small errors in those corrections result in very small overall errors. The choice of whether to use $H/2$ or $H/\sqrt{3}$ (or some other value) is therefore not one of critical import, and the selection can be made on the basis of other criteria. The slothful author uses the simpler value of $H/2$ almost all the time, since it minimizes another kind of energy, his own.

Noise Figure Measurement

1.0 Introduction

One of the most important performance metrics for low-level amplifiers is noise figure or noise factor. The two terms are used interchangeably in the literature. In this text, we adopt the following arbitrary convention: We will denote noise figure by *NF*, and define it as the decibel version of the noise factor, *F*. We will be somewhat sloppy about using the terms (reflecting common usage), but context should make clear whether or not the decibel version is being discussed.

The definition of noise factor now in use was first formally proposed by Harald Friis[1] of Bell Labs. At its core, the definition involves signal-to-noise ratios (SNRs):

$$F \equiv \frac{SNR_i}{SNR_o} \tag{1}$$

This definition shows that *F* is the factor by which an amplifier degrades the signal-to-noise ratio of the input signal. As such, it is never smaller than unity. As simple and straightforward as the definition appears to be, numerous subtleties are buried in it, and the definition is incomplete as presented. Accurate measurement of noise figure depends on a full appreciation of all of these subtleties, and an understanding of how to identify and correct sources of measurement error. As we'll soon see, automated noise figure instruments do not eliminate the need for a knowledgeable operator. As has been noted, "Automated equipment merely lets you produce more wrong answers per unit time." The purpose of this chapter is to reduce the rate of erroneous answer generation.

2.0 Basic Definitions and Noise Measurement Theory

One important subtlety concerns the temperature at which the measurement of noise figure is made. Specifically, the temperature of the source has a profound effect on the noise figure. Intuitively, this temperature dependence may be understood as follows: The device under test (DUT) generates its own internal noise, independent of the source temperature. If the latter is very low, then the source noise will be correspondingly low, so the noise added by the DUT will have a comparatively greater effect. The measured noise figure will thus be higher than if the source were hotter. Because of this sensitivity, a meaningful comparison of noise figures requires that the measurements be made at a standard temperature. Friis proposed a reference temperature, denoted T_0, of 290 kelvins (about 62°F or 17°C), a temperature which is considerably cooler than the interior of most laboratories. An oft-cited reason for this choice is the approximate equality of this temperature with that commonly seen by antennas used in terrestrial wireless communications. However, a

1. "Noise Figures of Radio Receivers," *Proc. of the IRE*, July 1944, pp. 419-422.

stronger motivation for its selection is simply that kT_0 is then 4.00 x 10^{-21}J, a round number with undeniable appeal in an era of slide rule computation, particularly to an eminently practical engineer like Friis.

The final statement on standard conditions, made by a committee of the Institute of Radio Engineers (a forerunner of the IEEE), is that the noise figure measurement is to be made with a source whose available noise power is the same as that of an input termination whose temperature is 290K. Recall that available power is defined as the power that *could* be delivered to a (conjugately) matched load. Hence, even if the source does not in fact happen to drive a matched load, the power remains *available*. Available power is precisely what the words imply: a potential power, independent of the actual load. Confusion about this definition is all too common, and can lead to serious errors, as will be made clear later in this chapter.

A second consideration is that determining input and output signal-to-noise ratios is by no means trivial. Since noise figure is an intrinsic property of the DUT alone (assuming linearity), and therefore not of how you drive the DUT, it should be possible to devise a measurement that does not involve the use of an explicit signal. To do so, it is helpful to note that the noise appearing at the output of the DUT results from two contributions. One is the amplified available source noise power (with the source at T_0 = 290K), which has a value

$$N_{os} = kT_0BG_{av}, \tag{2}$$

where B is the noise (brickwall) bandwidth and G_{av} is the available power gain of the DUT.

The other component of output noise is simply the noise added by the DUT itself. We call this noise contribution N_a. The total available output noise power is therefore

$$N_1 = kT_0BG_{av} + N_a. \tag{3}$$

Now let's revisit the noise figure definition of Eqn. 1:

$$F \equiv \frac{SNR_i}{SNR_o} = \frac{S_i/N_i}{S_o/N_o}. \tag{4}$$

Interpreting all quantities as available powers, the ratio of output signal S_o to input signal S_i is the available gain, G_{av}. The available input noise power is simply kT_0B, and the available output noise power is N_1 as defined in Eqn. 3. So we may write

$$F = \frac{S_i/N_i}{S_o/N_o} = \frac{1}{G_{av}}\left(\frac{N_o}{N_i}\right) = \frac{1}{G_{av}}\left(\frac{N_1}{N_i}\right) = \frac{N_1}{N_{os}} = \frac{kT_0BG_{av}+N_a}{kT_0BG_{av}}. \tag{5}$$

The last expression on the right,

$$F = \frac{kT_0BG_{av} + N_a}{kT_0BG_{av}}, \tag{6}$$

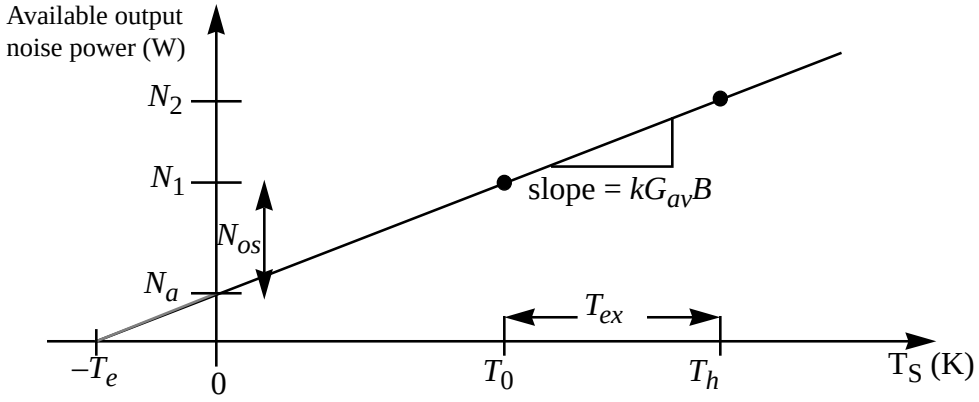
is the definition officially adopted by the IRE. It initially appears more attractive as a basis for measurement than Eqn. 1 because it contains no terms related to an explicit input or output signal. Using Eqn. 6, measurement of noise figure reduces to the measurement of noise, available gain and bandwidth. Unfortunately, there are still serious practical difficulties associated with trying to base a measurement directly on this equation. In particular, it is not easy to measure the product of the effective noise bandwidth and available gain, BG_{av}, with high accuracy. The experimental difficulties are best appreciated after comparing the various noise measurement methods discussed in Section 6.0.

One of these alternative noise figure evaluation methods, which is implemented in commercial instruments such as the HP8970A, cleverly sidesteps the need to measure gain-bandwidth by employing a ratio of noise measurements performed at two different source temperatures. As a general philosophy, it is always advantageous to replace absolute measurements with ratiometric ones wherever dimensional considerations permit it. Fortunately noise factor is a dimensionless quantity, so a purely ratiometric measurement is possible. Gain-bandwidth product is not dimensionless, so measuring it should not be fundamentally necessary here.

The basis for the ratiometric technique is that the use of a hot source increases the component of output noise due to the source, without changing the noise added by the DUT. If the ratio of the source temperatures is accurately known, then measuring the output noise powers under the hot and cold conditions permits us to solve for the noise added by the DUT and, hence, compute the noise figure.

The following plot of output noise power as a function of source temperature illustrates how such a ratiometric measurement solves our problem:

FIGURE 1. Output noise power vs. source temperature



Comparing features of this drawing with Eqn. 6, note that the slope and y-intercept tell us everything we need to compute F:

$$F = \frac{kT_0BG_{av} + N_a}{kT_0BG_{av}} = 1 + \frac{N_a}{kT_0BG_{av}} = 1 + \frac{\text{y-intercept}}{(T_0)\ (\text{slope})}\ . \tag{7}$$

Clearly, the need to measure gain-bandwidth has disappeared because two points determine a line. Despite the straightforward nature of this observation, engineers have devised a surprising number of different ways to use noise data from two points to determine noise figure. Just keep in mind that underlying the seeming complexity in what follows is the extremely simple geometric picture of Figure 1.

If we make a noise power measurement at a source temperature, T_h, that is above the reference temperature by an amount T_{ex}, then the available output noise power becomes

$$N_2 = kBG_{av}T_h + N_a = kBG_{av}(T_0 + T_{ex}) + N_a. \tag{8}$$

Combining the hot measurement with the one at T_0 (Eqn. 3), a little algebra allows us to find that the noise factor may be expressed as

$$F = \frac{T_{ex}/T_0}{\frac{N_2}{N_1} - 1}. \tag{9}$$

The ratio N_2/N_1 is often called the "*Y* factor" in the literature (why? because it comes after *X*...). Figure 1 shows a cold temperature equal to the reference temperature, T_0, but it should be clear that any temperature other than T_h could be used to figure out the slope and intercept of the line. More generally, if the cold temperature T_c is not T_0, the numerator changes, so that the noise factor is

$$F = \frac{\frac{T_{ex}}{T_0} - Y\left(\frac{T_c}{T_0} - 1\right)}{Y - 1}. \tag{10}$$

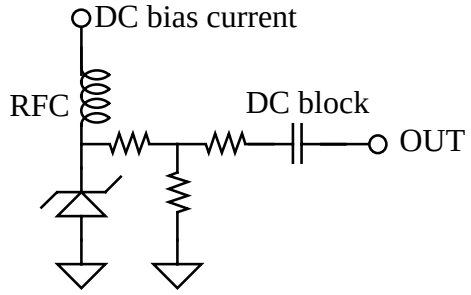
The ratio T_{ex}/T_0 is a property of the noise source, and is information (almost) supplied by the manufacturer. The qualifier "almost" applies because the manufacturer actually specifies a slightly different quantity called the *excess noise ratio* (ENR), which is defined as the ratio of noise powers actually delivered to a 50Ω load (or occasionally some other standard impedance level). However the ratio T_{ex}/T_0 results from a consideration of *available* powers (as does the *Y* factor). The two ratios are equivalent only in the special case where the noise source happens to have an impedance of precisely 50Ω. Despite the best efforts of manufacturers, this condition is not perfectly satisfied in practice, so substituting ENR for T_{ex}/T_0 is one (generally small) potential source of error. Because it is much easier to determine ENR, however, that's what the manufacturer measures and reports.

In the "old days," actual hot and cold sources were used, commonly with resistors at 77K (the boiling point of liquid nitrogen) and 373K (the boiling point of water). Clearly, the greater the temperature difference, the more accurately we can compute the slope and

intercept, for a given uncertainty in the power measurement. A limitation on the hot side is the difficulty of accurately determining or controlling the temperature. And the higher the temperature, the more significant the problems of materials properties (e.g., melting).

Nowadays, it is common to use noise diodes (see the chapter on RF diodes) which can produce the noise of an exceptionally hot source (e.g., 10,000K, higher than the melting point of any known metal) while remaining at room temperature. The same diode can provide the cold reference as well, simply by turning it off, causing an internal resistive matching network to provide an available noise power that corresponds to the ambient temperature (RF choke RFC is simply an inductor large enough to be considered an open-circuit at all frequencies of interest):

FIGURE 2. Typical noise diode

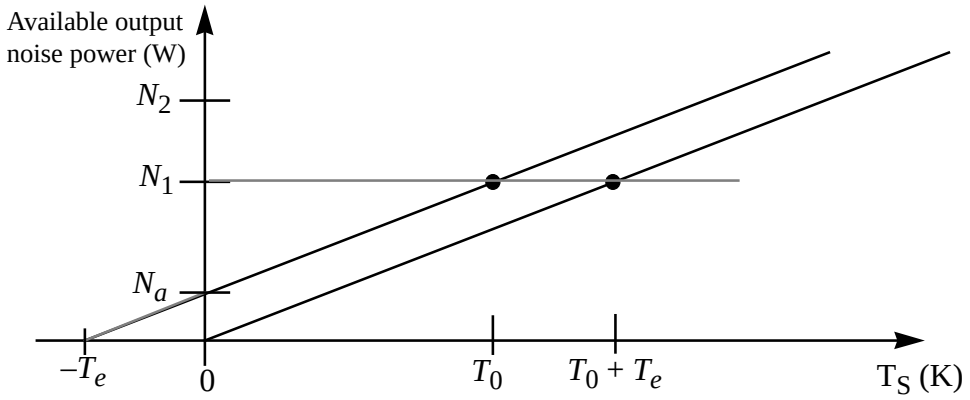


One drawback is that, unlike true hot and cold resistors, such diodes are not fundamental standards; their hot noise cannot be computed from first principles. Since ENR must be known to great accuracy to be useful, it is usually traceable to a primary noise standard maintained by national laboratories, such as the NIST (National Institute for Standards and Technology, formerly the National Bureau of Standards). This traceability accounts in part for the relatively high cost of noise diodes.

3.0 Noise Temperature

Noise temperature, T_e, is an alternative figure of merit used in place of noise figure in some cases. As seen in Figure 1, noise temperature is (minus) the extrapolated intercept of the noise power curve with the temperature axis. An intuitively appealing meaning of noise temperature can be extracted by translating the noise power curve to the right by a temperature equal to the noise temperature:

FIGURE 3. Noise temperature



The translated curve is that of a noiseless amplifier (because the noise at zero source temperature is zero) with the same slope (= available gain-bandwidth product, times *k*) as the original amplifier. As can be seen, this noiseless amplifier produces an available output noise power equal to the available output noise of the original amplifier, if the source is now heated to a temperature $T_0 + T_e$. The increase in available output noise power due to the hotter source is precisely equal to the available noise (N_a) added by the original DUT:

$$N_a = kT_eBG_{av}. \tag{11}$$

Noise temperature is used most often in satellite communications systems for several reasons. One is that objects in the sky generally don't have an effective temperature anywhere near 290K, so choosing such a reference temperature has a weaker physical justification. The other is that space communication systems generally have exceptionally low noise figures, and noise temperature is a higher resolution measure of very low noise figure values. The following table compares noise figure, noise factor and noise temperature over a range generally considered very low noise:

TABLE 1. Comparison of noise figure, noise factor and noise temperature

NF **(dB)**	***F***	T_e **(kelvins)**
0.5	1.122	35.4
0.6	1.148	43.0
0.7	1.175	50.7
0.8	1.202	58.7
0.9	1.230	66.8
1.0	1.259	75.1
1.1	1.288	83.6
1.2	1.318	92.3

It is sometimes helpful to note that, in the very low noise figure regime (e.g., below about 1dB), the noise figure in dB is approximately the noise temperature divided by 70-75. Stated alternatively, each tenth of a dB corresponds to roughly 7-7.5K.

To relate noise temperature and noise factor, return again to the official IRE noise figure definition:

$$F \equiv \frac{N_1}{N_{os}} = \frac{kT_0BG_{av} + N_a}{kT_0BG_{av}}. \tag{12}$$

Substituting Eqn. 11 for N_a yields

$$F = \frac{kT_0BG_{av} + N_a}{kT_0BG_{av}} = \frac{kT_0BG_{av} + kT_eBG_{av}}{kT_0BG_{av}}, \tag{13}$$

which simplifies to

$$F = 1 + \frac{T_e}{290}. \tag{14}$$

If the noise added by the DUT equals the noise power of the source, the noise figure will be 3dB, corresponding to a noise temperature of 290K. Many LNAs with effective noise temperatures well below 100K (corresponding to noise figures below 1.3dB) are commercially available.

The noise temperature may be found indirectly by relating Eqn. 14 to Eqn. 9, or directly from the hot and cold noise measurements of Figure 1. Pursuing the latter strategy, we may write

$$N_2 = kT_hBG_{av} + N_a = k(T_e + T_h)BG_{av} \tag{15}$$

and

$$N_1 = kT_cBG_{av} + N_a = k(T_e + T_c)BG_{av}, \tag{16}$$

so that

$$Y = \frac{N_2}{N_1} = \frac{k(T_e + T_h)BG_{av}}{k(T_e + T_c)BG_{av}}. \tag{17}$$

Solving for T_e yields

$$T_e = \frac{T_h - YT_c}{Y - 1}. \tag{18}$$

Another reason that noise temperature is used in some contexts is that the quantity $F - 1$ recurs frequently in certain calculations (particularly of cascaded noise figure, as we shall see in Section 4.0). By rearranging Eqn. 14, it's clear that noise temperature T_e is proportional to $F - 1$, so its use simplifies such calculations.

Because both noise figure and noise temperature fully convey the information of the other (as implied by Eqn. 14, for example), you may use either. The choice of which to use is made largely on the basis of culture and convenience.

3.1 Spot noise figure

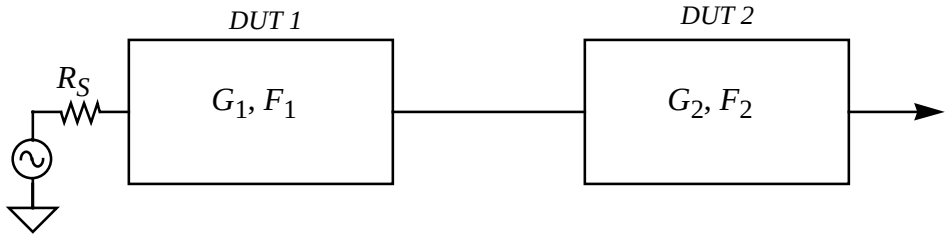
In many cases, one is interested in the noise performance of an amplifier as a function of frequency. In those situations, the measurement bandwidth is restricted to some known value (e.g, 1MHz) and the noise figure for that bandwidth is reported at a specific frequency. Since the parameter is a noise figure measured in a narrow band centered around a specific spot, it is known as the spot noise figure. The noise figures most often reported in the literature are usually spot noise figures.

4.0 Friis' Formula for the Noise Figure of Cascaded Systems

Computing the noise figure of a cascade of systems is often carried out incorrectly. Once again, the problem is a failure to appreciate certain subtleties. One difficulty is that individual noise figures do not combine in any simple way to yield the overall cascaded noise figure. Another is that each stage may see a different source impedance, and the noise figure of each stage must be computed with respect to that impedance. To understand these and other issues in detail, we now derive the correct equation for the cascaded noise figure, called Friis' formula.

Consider a noisy system that is driven by yet another noisy system:

FIGURE 4. Cascaded systems



The first stage has a noise factor F_1 and available power gain G_1 measured with R_S as source resistance. The second stage has an available power gain G_2 and a noise factor F_2 when these quantities are measured *with the output impedance of the previous stage as source resistance*. If there were additional stages, the available gain and noise figure of each one would be determined using the output impedance of the preceding stage as the source resistance. A common error is to use R_S as the source impedance for all stages, but this choice is correct only if the output impedances happen to be R_S.

The easiest way to derive Friis' formula is to make use of the concept of noise temperature. Because the available output noise power added by each DUT is kT_eBG_{av}, the available noise power at the output of the first DUT is

$$N_{o1} = kT_sBG_{av1} + kT_{e1}BG_{av1} = k(T_s + T_{e1})BG_{av1}. \quad (19)$$

The second stage takes this noise, amplifies it, and adds to it another kT_eBG_{av} of its own:

$$N_{o2} = k(T_s + T_{e1})BG_{av1}G_{av2} + kT_{e2}BG_{av2}. \quad (20)$$

We could just as well regard the overall system as a single amplifier with available gain $G_{av1}G_{av2}$, driven by a source R_s. Hence, we may also write

$$N_{o2} = k(T_s + T_{e12})BG_{av1}G_{av2}, \quad (21)$$

where T_{e12} is the overall noise temperature of the cascade. Equating Eqn. 20 and Eqn. 21 yields

$$T_{e12} = T_{e1} + \frac{T_{e2}}{G_{av1}}. \quad (22)$$

The overall noise temperature is therefore the noise temperature of the first stage, plus the input-referred noise temperature of the second stage. This formula reflects the understanding that the signal boost provided by the first stage diminishes the effect of noise of subsequent stages. Clearly, Eqn. 22 can be extended to an arbitrary number of stages, yielding one form of Friis' formula:

$$T_{e12} = T_{e1} + \frac{T_{e2}}{G_{av1}} + \frac{T_{e3}}{G_{av1}G_{av2}} + \ldots. \quad (23)$$

An alternative expression in terms of noise factors is readily derived by using Eqn. 14 to relate noise temperature and noise factor:

$$F_{12} = F_1 + \frac{F_2 - 1}{G_{av1}} + \frac{F_3 - 1}{G_{av1}G_{av2}} + \ldots. \quad (24)$$

From inspection of the last two equations, we see that the expression for cascaded noise temperature is somewhat simpler (none of those pesky –1 terms to clutter up the equation). The noise temperature contributed by the *n*th stage can be computed simply by dividing through by the product of the available gains of the $(n - 1)$ stages preceding it. For this reason, the noise temperature formulation is frequently favored when considering cascaded systems.

5.0 Noise Measure

From Friis' formula, we see that if an amplifier has good noise figure but low gain, suppression of noise from subsequent stages is poor. Unfortunately, classical noise optimization design methods sometimes lead to an "optimum" amplifier design with precisely this combination of characteristics. Because both the noise figure and gain of an amplifier are

important in general, another figure of merit known as *noise measure* is sometimes used to guide engineers toward a balanced design. Its formal definition initially seems to combine these two quantities in a puzzling way:

$$M \equiv \frac{F-1}{1-\frac{1}{G_{av}}} \tag{25}$$

The rationale for this definition becomes clear when we examine Friis' formula for the special case of an infinite cascade of identical amplifiers:

$$F_{tot} = F + \frac{F-1}{G_{av}} + \frac{F-1}{G_{av}^2} + \ldots, \tag{26}$$

which ultimately simplifies to

$$F_{tot} = 1 + \frac{F-1}{1-\frac{1}{G_{av}}} = 1 + M. \tag{27}$$

Therefore, this definition of noise measure is actually the normalized noise temperature of the infinite cascade:

$$T_{e,\,tot} = (F_{tot} - 1)\,T_0 = MT_0 \Rightarrow M = \frac{T_{e,\,tot}}{T_0}. \tag{28}$$

Just to keep you on your toes, though, noise measure is defined in some references as F_{tot}, rather than as $F_{tot} - 1$. Be sure to identify which definition is being used, as the difference can introduce considerable error for low noise systems. Finally, note that this definition of noise measure has no particular relationship to the definition of noise measure for negative resistance devices, such as Gunn and tunnel diodes (see chapter on RF diodes).

6.0 Typical Noise Figure Instrumentation

Having derived multiple expressions for noise figure, we're now in a position to examine several different methods for carrying out an actual measurement. As usual, we start with a little history, partly for entertainment, but partly because methods that were used long ago tend to be ones that hobbyists can implement economically today.

6.1 The (good?) old days

From Figure 1 we see that measuring noise figure is equivalent to determining the equation of the noise power-vs.-source temperature line. Measuring two points along the line is sufficient, but so is knowing a single point and the line's slope. The former method is the

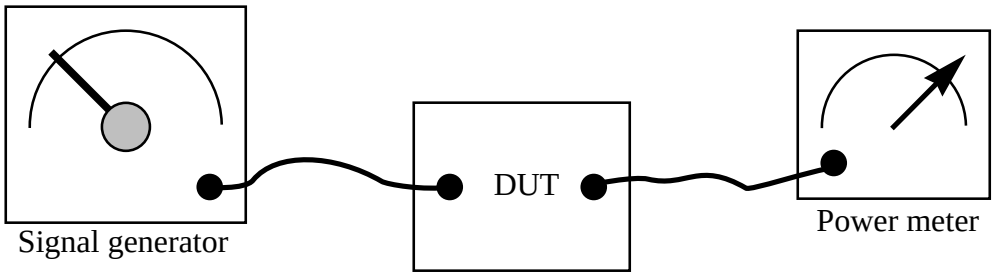
modern way, but it is worthwhile discussing the latter. Even though it poses nontrivial experimental challenges, the equipment required is within the reach of most RF hobbyists, so a description of this technique merits inclusion here.

Prior to the development of calibrated hot and cold sources, the only noise source available was one at room temperature. With that limitation, one can determine the available output noise only at that one (perhaps inaccurately known and poorly controlled) temperature. So immediately, we see one error source: the noise source is probably at a temperature higher than 290K, unless the laboratory is also used for storing beer. Even so, this error source is usually not the dominant one.

The tricky part is that of determining the slope of the line, $kG_{av}B$. Boltzmann's constant is pretty solid, but measuring the product of available power gain and noise bandwidth (which generally does not equal the –3dB bandwidth) is fraught with difficulty. The experimental setup for doing so is straightforward in principle; it's just the practice that's hard.

To measure $G_{av}B$, simply connect a signal generator to the DUT and sweep the frequency to plot the power gain-vs.-frequency curve:[2]

FIGURE 5. Signal generator method for noise figure measurement



In most cases, no provisions are made to ensure a conjugate match (because it is exceedingly tedious to do so at each of many test frequencies), so the power gain that is measured differs from the available gain, leading to potential errors. The power frequency response curve is integrated (e.g., graphically, or by measuring the –3dB bandwidth and multiplying by some fudge factor between 1 and 1.57) to find the product GB.

To complete the experiment, the output power N_1 is measured with the noise source (e.g., a simple resistor of value R_s) connected to the input. The noise factor is then

$$F \equiv \frac{N_1}{N_{os}} = \frac{N_1}{kT_0BG_{av}}. \tag{29}$$

This measurement method requires simple apparatus: a signal generator, calibrated power meter (or oscilloscope; see the appendix for eyeball methods of estimating noise) and a resistor (which might be provided by simply turning off the generator). Figuring out the

2. If the signal generator's output is not constant over the band, it is necessary to measure its output to perform a proper gain calculation. Failure to do so is a common source of error.

gain-bandwidth product from the measured frequency response is rather labor-intensive, but if you don't want to have to choose between buying a car and buying an automated noise figure meter, the traditional method is the best choice. That said, it is quite difficult to reduce noise figure uncertainties below about 1-2dB with this method, so characterization of very low noise amplifiers with this technique is generally out of the question, practically speaking.

Another issue is that the measurement time per frequency point is large, so that it is cumbersome to make real-time evaluations of tweaks made to improve noise figure. It takes patience to use the signal generator method.

There is one case (at least, this is the only one the author can think of) where the signal generator method is favored, however. Consider the problem of measuring accurately the noise figure of an exceptionally noisy system. In particular, suppose that the DUT is so noisy that the noise temperature greatly exceeds the reference temperature. In this case, it is possible for the output noise powers under the hot and cold conditions to be rather similar, leading to a Y factor close to unity. Because the formula for noise factor with a hot/cold measurement method contains a term $(Y - 1)$ in the denominator, the measurement can be quite sensitive to small errors in Y when Y is nearly unity. The signal generator method, on the other hand, does not suffer from this sensitivity because it does not infer slope from measuring noise at two temperatures; there are no subtractions along the way. So it may be said that, for low noise amplifiers, the hot/cold method is better, and for extremely noisy systems, the signal generator method may be better.

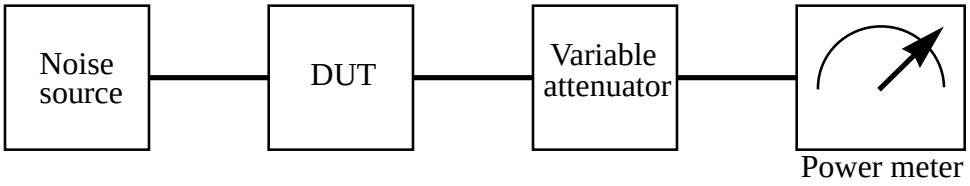
6.2 On to the modern era...

When a calibrated hot noise source is available, it becomes considerably easier to make accurate noise figure determinations. As mentioned previously, early sources used actual resistors heated or cooled to easily determined or controlled temperatures, such as the boiling points of water and liquid nitrogen. A hot source at the temperature of boiling water is entirely feasible for the home experimenter (just be careful not to burn yourself or start a fire), and highly accurate as long as the water is reasonably pure and corrections are made for boiling point shifts with altitude.

Many commercial cold loads operate at 77K, but not many hobbyists happen to have a Dewar full of LN_2 about the house. Perhaps a more practical choice for the weekend experimenter is to use a room temperature cold source, but accurate measurements demand knowledge of the actual room temperature. A modest improvement is possible by using ice in water to provide a 273K cold temperature. If you have access to acetone and dry ice, an equilibrium mixture of those two substances will have a temperature of about 203K (–70°C). However, acetone is quite flammable, so if you do choose this mixture, be sure to observe all appropriate safety precautions (in particular, keep the acetone well away from whatever makes the hot source hot). Also, it is important to keep in mind that acetone is a powerful solvent for most plastics, so you can't put the mixture in a styrofoam cup! Additionally, its fumes are toxic as well as flammable, so use only in a well ventilated room. Except for the possibility of death by fire, asphyxiation or cancer, this mixture is ideal for realizing the cold source at home.

Once you have both a hot and cold source, there are several measurement options from which to choose. One method, called the "*Y*-factor method" for reasons that will become clear, avoids the need for a calibrated power meter, replacing it instead with a more easily realized calibrated adjustable attenuator:

FIGURE 6. *Y*-factor measurement technique (simplified)



This measurement technique relies on the fact that the ratio of output powers with the hot and cold source (= *Y*), plus knowledge of the hot and cold temperatures, is sufficient to compute noise figure. To carry out a measurement with this method, set the attenuation factor to unity, connect the cold load, and note the output power reading on the meter. The absolute value is completely unimportant. Then connect the hot load and adjust the attenuator until you obtain the same power reading as before. Since the attenuation factor is therefore the value that reduces a power N_2 to a value N_1, the attenuation factor is precisely equal to *Y*. The noise factor is then computed from Eqn. 10:

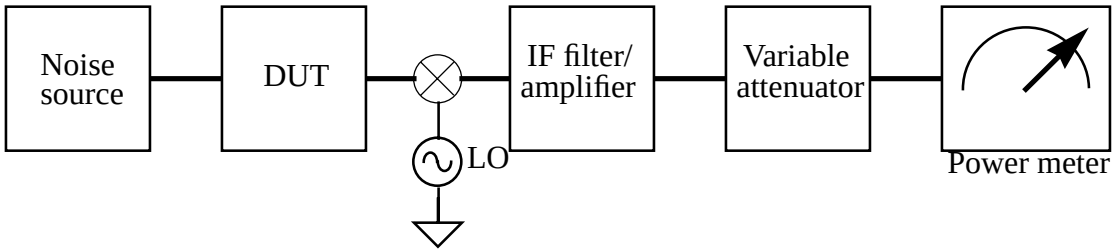
$$F = \frac{\frac{T_{ex}}{T_0} - Y\left(\frac{T_c}{T_0} - 1\right)}{Y - 1}. \tag{30}$$

The accuracy achieved depends on the accuracy of the *Y* factor determination, as well as on the knowledge of the hot and cold temperatures. With assiduous attention to controlling all error sources, this technique can provide accuracies that are on a par with what can be achieved with commercial instrumentation. The tradeoff is again one of time per measurement.

Actual *Y* factor noise determinations are usually carried out with a slightly different configuration to permit measurement of spot noise figure as a function of frequency, rather than a gross noise figure over the entire bandwidth of the amplifier. The typical setup modifies the one shown in Figure 6 by adding a mixer, local oscillator and intermediate-frequency (IF) amplifier, just as in a superheterodyne receiver:[3]

3. An LO, mixer and IF amplifier can also be added to the setup of Figure 5 to improve the signal generator method, readily permitting evaluation of spot noise figure with that system. The noise bandwidth of the IF filter determines the width of the spot, and its value needs to be known to calculate spot noise figure correctly.

FIGURE 7. More typical *Y*-factor measurement setup



Here, the LO frequency is swept to sample the noise from the DUT at different frequencies. The IF amplifier/filter combination ensures that this noise is measured over some narrow, controlled bandwidth centered about the frequency determined by the LO setting. In many implementations, a filter is additionally interposed between the DUT and the instrumentation to limit bandwidth (perhaps to attenuate image response, for example).

If a spectrum analyzer is used as the power meter in any of these methods, it is necessary to precede it with a high-gain, low-noise preamplifier because spectrum analyzers generally have rather high noise figures (e.g., 30dB), as a result of design trade-offs made in favor of good large signal linearity. The gain of the preamp must be large compared with the noise figure of the analyzer in order to effect a substantial reduction in *NF*. The overall noise figure will then be close to that of the preamp alone. Even so, the noise figure of the preamp-spectrum analyzer combination will generally remain high enough that it cannot be ignored, and Friis' formula for cascaded noise figure should be used to correct the measured values. As a concrete numerical example, assume that the preamp has a noise figure and available power gain of 3dB and 40dB, respectively, and that the analyzer has a noise figure of 30dB. The combination has a noise factor given by Friis' formula,

$$F = 2 + \frac{1000 - 1}{10^4} \approx 2.1, \tag{31}$$

or 3.2dB, a large improvement over 30dB and now only slightly greater than the preamp's inherent noise figure. Friis' formula will ultimately be used again, once the DUT is connected to the combination, with 3.2dB now considered the second stage's noise figure, and the DUT's available gain used in the denominator. If the noise figure of the preamp is not known, connect the hot and cold sources directly to its input, and make a measurement of noise figure. Once the combination has been characterized in this manner, it may be used to determine the noise figure of the DUT.

When making spot noise figure measurements with the spectrum analyzer, set the analyzer's resolution bandwidth equal to the desired width of the spot. Choose a video bandwidth (much) narrower than the resolution bandwidth to reduce noise in the displayed data (recall that video bandwidth controls the averaging of the *output* signal, after the detector).

The basic arrangement shown in Figure 6 is also quite close to what lies at the core of most modern automatic noise figure instruments. The HP8970, for example, contains all of those components, with the addition of a filter and preamplifier (as in the spectrum ana-

lyzer example) and a collection of attenuators at the input and output. With these additional elements, the instrument is able to measure (and correct for) insertion gain (loss) of the DUT and fixturing during the noise figure measurement. Additionally, the noise figure of the meter circuitry must also be known in order to complete an accurate noise figure measurement. What follows is a typical sequence of operations for making a noise figure measurement with a commercial instrument (specifically, the 8970B).

1) Read off the ENR calibration values for the hot/cold noise source, and enter those numbers into the instrument's memory. The 8970B has a list of the common calibration frequencies already in ROM, so the user normally only has to enter the ENR values.

2) Select the start frequency, stop frequency and frequency increment (step size).

3) Connect the noise source to the instrument in order to permit measurement of the meter's noise figure, set frequency to the desired value, and press the "calibrate" key to initiate the calibration sequence. The meter successively activates and deactivates the noise source to compute the meter's hot and cold noise powers:

$$P_{hm} = k(T_h + T_{em})BG_m \tag{32}$$

$$P_{cm} = k(T_c + T_{em})BG_m. \tag{33}$$

The ratio of these two powers is completely insensitive to gain-bandwidth product and has only the noise temperature of the meter as an unknown:

$$\frac{P_{hm}}{P_{cm}} = \frac{T_h + T_{em}}{T_c + T_{em}} \Rightarrow T_{em} = \frac{P_{cm}T_h - P_{hm}T_c}{P_{hm} - P_{cm}}. \tag{34}$$

The 8970 allows the results of several calibration runs to be averaged. The number of runs is controlled with the "increase" key. Hold it down until the desired number of runs is displayed. This step precedes activation of the "calibrate" mode.

4) Insert the DUT between the noise source and the instrument and select "noise figure and gain." To the maximum practical extent, avoid cables. The shorter the fixturing, the better, to minimize pre-DUT loss (and thus any errors introduced by uncertainties in its subsequent subtraction). The instrument then measures the hot and cold powers of the cascade (DUT + meter):

$$P_{h,tot} = k(T_h + T_{e,tot})BG_mG_{DUT} \tag{35}$$

$$P_{c,tot} = k(T_c + T_{e,tot})BG_mG_{DUT}. \tag{36}$$

The ratio of these two powers is also insensitive to gain-bandwidth product and has only the noise temperature of the DUT/meter combination as an unknown:

$$\frac{P_{h,\,tot}}{P_{c,\,tot}} = \frac{T_h + T_{e,\,tot}}{T_c + T_{e,\,tot}} \Rightarrow T_{e,\,tot} = \frac{P_{c,\,tot}T_h - P_{h,\,tot}T_c}{P_{h,\,tot} - P_{c,\,tot}} \tag{37}$$

The ratio of the differences of noise powers enables computation of the DUT's gain:

$$G_{DUT} = \frac{P_{h,\,tot} - P_{c,\,tot}}{P_{hm} - P_{cm}}. \tag{38}$$

The gain of the meter has dropped out completely, so its value is theoretically irrelevant. Having computed the gain of the DUT, the noise temperature of the meter, and the noise temperature of the meter and DUT combination, Friis' formula can be used to solve for the noise figure of the DUT alone:

$$T_{e,\,tot} = T_{DUT} + \frac{T_{em}}{G_{DUT}} \Rightarrow T_{DUT} = T_{e,\,tot} - \frac{T_{em}}{G_{DUT}}. \tag{39}$$

Note that the resulting calculation is correct only if G_{DUT} is equal to the available gain. Mismatches may make these unequal and thereby introduce error.

The 8970 also allows the user to enter the cold temperature. The default is 296.5K, which is closer to typical room temperatures.

A separate measurement of fixturing loss (e.g., with a network analyzer) enables correction for any pre-DUT fixturing attenuation. Most instruments allow the user to enter loss values (via the "loss compensation" feature of the 8970, for example), and automatically perform the subtraction of the loss factor. The instrument converts noise temperature into noise figure, and displays both *NF* and G_{DUT}. It takes you much longer to read this description than it does for the instrument to carry out the measurement.

7.0 Error Sources

There are several ways in which noise figure measurements can go awry. Understanding what these are is a key to making accurate measurements. What follows is a short list of common problems, mistakes and their fixes.

7.1 External noise

More than occasionally, external interference couples into the test setup. This interference can be noise radiated by RF sources ranging from TV and radio, to digital equipment (particularly computers and their monitors). Noise figure measurements are best carried out in a shielded screen room to prevent this interference from injecting into the system. If this option is not available, the next best choice is to make a spot noise measurement at a frequency removed from the interference, assuming that it is narrowband enough to enable this strategy. Many noise figure measurement systems provide for an oscilloscope connec-

tion to monitor the spectrum (generally at the output of the noise figure meter's IF stage). If such an output is not available, a normal spectrum analyzer may be used instead. With the aid of a monitor, discrete peaks caused by interference can be identified quite easily, and the measurement frequency moved appropriately away from the interference.

7.2 Fixturing loss

Fixturing anomalies are an endless source of errors. For example, a proper measurement of noise figure requires accurate characterization of any loss (e.g., from cable attenuation) that precedes the DUT proper. This loss (in dB) is subtracted from the measured overall *NF* to yield the DUT's true *NF*. If the loss is large, however, the uncertainty in the final answer can be considerable because the instrument will have subtracted two nearly equal numbers. For example, suppose that the pre-DUT fixturing power loss is 20dB (this is frighteningly large), and the DUT itself has a 2dB noise figure. The noise figure meter will measure a 22dB noise figure, but within some error (say, optimistically, 0.5dB). Assume for now that the error results in a composite measured *NF* of 21.5dB. A separate measurement of the fixturing loss might have a similar uncertainty of 0.5dB; suppose we measure 20.5dB in this case. After subtraction, we compute a DUT *NF* of 1dB, instead of the correct value of 2dB, a huge error. In fact, for amplifiers with very low noise figure, it is entirely possible to compute negative values! Therefore, be suspicious of noise figure measurements in which a large attenuation has been mathematically removed. As a general rule, it is desirable to limit any such pre-DUT attenuation to values smaller than the anticipated noise figure. The lower this loss, the better.

7.3 Second stage contribution

Another common error is a failure to take into account the noise of stages that follow the DUT (the "second stage contribution"). A related consideration is that all commercial noise figure meters assume that the measured DUT gain is the same as the available gain. If the DUT has a large output impedance mismatch with that of the noise figure meter's input port, this assumption will be a poor one, and the calculation of the second stage contribution will be in error, as can be seen from Eqn. 39.

Impedance mismatch between the output of the noise source and the input of the DUT is also a concern. Reflections off of the DUT input travel back to the noise source where any mismatch there causes a re-reflection back toward the DUT. The superposition of the incident power and this reflected power can cause the noise power from the source to differ from what it would be with a matched load. Complicating the situation is that the noise source may have a different impedance in the hot and cold modes, compounding the error.

Correction for all of these errors requires knowledge of all three of the mismatches, as can be seen from the following formula:[4]

4. *Fundamentals of RF and Microwave Noise Figure Measurements*, HP Application Note 57-1, July 1983.

$$K_G = \frac{(1-|\Gamma_s|^2)\,|1-\Gamma_1\Gamma_2|^2}{(1-|\Gamma_2|^2)\,|1-\Gamma_1\Gamma_s|^2}. \quad (40)$$

Here, K_G is the factor by which the measured insertion gain should be multiplied in order to yield the correct value of available gain. The reflection coefficients are referred to various ports as follows: Γ_1 is defined looking into the input of the noise measurement instrumentation, Γ_2 into the output of the DUT, and Γ_s into the output of the noise source. Note that if these reflection coefficients are zero, K_G is unity. Note also that knowledge of the both the magnitude and phase of the reflection coefficients is necessary to perform the correction. If only the magnitudes of the reflection coefficient are known, the best one can do is bound the error. As a specific example of the latter, assume that the magnitudes of Γ_1, Γ_2, and Γ_s are 0.33, 0.33, and 0.11, respectively. Then the true available gain could be anywhere between about 0.95 and 1.3 times the measured insertion gain.

7.4 Noise source calibration uncertainty

Uncertainty in the ENR of the noise source is an additional error source. As stated earlier, noise diodes must be calibrated against a standard. Calibrations are never perfect, and noise diodes are not perfectly stable (although good ones are remarkably good). One may typically expect instrument-grade noise diodes (such as the popular HP346B) to possess uncertainties in ENR on the order of 0.1dB at low frequencies (e.g., 10MHz), increasing to perhaps 0.2dB at higher frequencies (e.g., 18GHz). The percentage error represented by these uncertainties gets progressively more significant as the noise figure of the DUT diminishes.

Because a noise diode's output is not perfectly constant over the operating frequency range, nor follows any other simple functional law, noise source calibrations are made at a number of discrete frequencies (10 or 20 is a typical number). In between the calibration points, you (or the noise figure meter) has to perform interpolations. The actual noise output may differ from the interpolated value, adding another error term.

7.5 Cold temperature $\neq T_0$

Yet another common problem is that the cold noise source temperature is rarely 290K. A diode noise source has a cold temperature equal to that of the ambient, and most laboratories are 4-5°C warmer than T_0. As a rough rule of thumb, the measured noise temperature is too low by one degree for each degree the noise source is above T_0. Thus it is typical to underestimate the noise figure of a DUT because of the warm laboratory problem. For more rigorous corrections, an accurate measurement of the cold temperature must be made, and Eqn. 10 used to compute the adjustment. This correction is most important in the case of very low noise figures.

7.6 Failure of linearity: diode detectors

The straight line of Figure 1 underlies both the definition and measurement of noise figure. If the device under test is nonlinear, noise figure can't be uniquely defined. A relevant example is that of diodes used as square-law detectors (frequently known as video detectors for historical reasons). In cases such as these, a different figure of merit is used to convey information about noise performance.

One such figure of merit is *tangential signal sensitivity* (TSS). Its original definition is a highly subjective evaluation of noise: An operator observes the noisy output of the detector on an oscilloscope in the absence of any signal and notes the position of the positive noise peaks. Then the signal is turned on, and the operator adjusts the amplitude until the negative-going noise peaks with signal present appear just to touch the positive-going noise peaks noted earlier with the signal absent. TSS is defined as the level of input signal that produces this condition. The problem with this definition is that noise, being random, has theoretically unbounded peaks. So, the operator has to make an arbitrary judgment when an equality of peaks occurs, and different operators may guess differently (the same operator may also make different determinations at different times). To eliminate this subjectivity, most diode manufacturers now define TSS as the available input signal power that causes an output power SNR of 8dB. A typical value of TSS for diodes might be –60dBm.

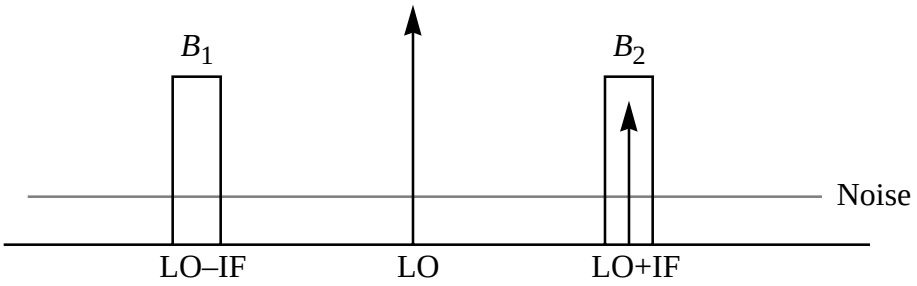
Another figure of merit is the *nominal detectable signal* (NDS) which is defined as the available input power that results in an output SNR of unity. Both TSS and NDS are generally functions of frequency and bias current, so these must be specified to make TSS and NDS values meaningful.

8.0 Special considerations for mixers

When the DUT is a mixer, there is a question of whether one should perform a single-sideband (SSB) or double sideband (DSB) noise figure measurement. In most cases, the SSB noise figure is the appropriate choice, as few communications systems transmit the same signal in both the main and image bands. The only two exceptions the author is aware of are direct-conversion (homodyne) receivers, in which the main signal occupies the same spectrum as its image, and in deep-space radiometry where noise (that of the universe) *is* the signal. Because DSB noise figure is lower by 3dB (assuming equal conversion gains for the two sidebands), "specmanship" games are all too frequent, and this figure is often reported instead of SSB.

To place the DSB-SSB issue on a firm foundation, consider that the IRE (now IEEE) noise figure definition has in its numerator all output noise, but has in the denominator only signal-related noise. If the signal is contained in only one sideband, then the relevant spectra appear roughly as follows:

FIGURE 8. Spectrum of SSB input to mixer



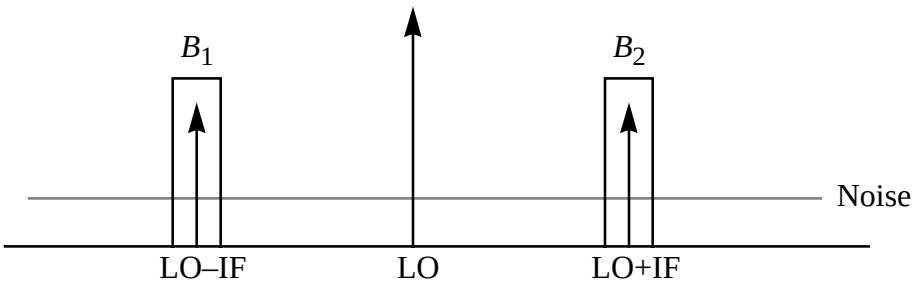
Here the signal exists only within bandwidth B_2. From this picture the correct definition of noise factor is:

$$F_{SSB} \equiv \frac{N_a + kT_0G_1B_1 + kT_0G_2B_2}{kT_0G_2B_2}. \tag{41}$$

Note that the formula allows for the possibility of unequal receiver bandwidths and unequal conversion gains for the two bands.

In the rarer DSB case, the desired signal resides in both bands:

FIGURE 9. Spectrum of DSB input to mixer



The corresponding noise factor is:

$$F_{DSB} \equiv \frac{N_a + kT_0G_1B_1 + kT_0G_2B_2}{kT_0G_1B_1 + kT_0G_2B_2}. \tag{42}$$

If the bandwidths are equal, and the conversion gains are equal, the DSB *NF* will be 3dB lower than the SSB value, as stated earlier. More generally, allowing for unequal conversion gains (but still assuming equal bandwidths),

$$F_{SSB} = F_{DSB}\left(1 + \frac{G_1}{G_2}\right). \tag{43}$$

In many cases, a mixer is preceded by an image-suppression filter. In this commonly occurring situation, it is appropriate to characterize the combination of the filter and mixer as a unit. Because the job of the filter is to produce unequal conversion gains to the two sidebands, there will no longer be a 3dB difference between the SSB and DSB *NF*.

Another important subtlety concerns the nature of terminations on the several mixer ports. Because a mixer has three ports – RF, IF and LO – misterminations on any of the ports can result in complicated reflections capable of corrupting measurements. A particularly common error is to terminate the IF port of a passive mixer in a load that is matched only at, say, the difference frequency, exhibiting a highly reactive impedance at the sum frequency. Even though we might only wish to use the difference component, the sum component nonetheless exists also. Reflections at the sum frequency can cause pathological behavior of both noise figure and conversion gain.

Finally, the gain and noise characteristics of mixers typically vary with LO power. For the measurements to be meaningful, then, the LO power must be specified. Preferably, the noise figure (and conversion gain) should be presented as a function of LO power over a range that spans practical values.

9.0 References

Various applications notes from Hewlett-Packard (now Agilent Technologies) are excellent sources of information about noise measurement. Some that are of particular interest include:

Accurate and Automatic Noise Figure Measurements, HP Application Note 64-3, June 1980.

Fundamentals of RF and Microwave Noise Figure Measurements, HP Application Note 57-1, July 1983.

Another good source of information is the documentation for the HP8970B noise figure meter, which describes in detail the theory underlying the operation of this instrument.

10.0 Appendix: Two Cheesy Eyeball Methods

For very quick assessments of relatively large amounts of noise, a crude measurement is sometimes acceptable. In those cases, an oscilloscope and your eyeball may be the only instruments you need. If we assume that the noise is Gaussian, then the peak-to-peak values very rarely exceed about 5-7 times the rms value. So, the level zero eyeball measurement is to connect the noisy DUT to the oscilloscope, make some judgment about what the displayed peak-to-peak value seems to be, then divide by about six to develop an estimate of the rms value.

This method is *very* crude, of course, and in no small measure because of the difficulty in determining what the “true” peak-to-peak value happens to be. The situation is further

complicated by the fact that the oscilloscope brightness setting affects what appear to be the peaks; the brighter the trace, the taller the apparent peaks. And, as with TSS, the same operator may also make significantly different determinations at different times, as a function of sleep deprivation, emotional state, and caffeine levels.

A clever extension of the eyeball technique removes much of this uncertainty by converting the measurement into a differential one.[5] Here, the noisy signal drives both channels of a dual-trace oscilloscope. With a sufficiently large initial position difference, there will be a dark band between these two traces. The operator adjusts the position controls until the dark band just disappears, with the two traces merging into a single blurry mess with a monotonically decreasing brightness from the center outward. Note that this description implies an independence of the result on the absolute intensity. The noisy signals are then removed, and the distance between the two baselines measured. The resulting value is twice the rms voltage to a good approximation. Absolute accuracies of about 1dB are possible with this simple method.

The basis of this technique is that a sum of two identical gaussian distributions has a maximally flat top when the two distributions are separated by exactly twice the rms value.

Because the eye is an imperfect judge of contrast, it is not possible to establish with infinite precision when the dark band disappears. When following the procedure as outlined, most people will perceive the band to have disappeared a little before it actually does. The error resulting from this uncertainty is on the order of 1dB for most people. Thus, perhaps 0.5dB should be subtracted from the measurement if you are very fussy. An alternative is to measure the noise two different ways, one using the procedure given, and another with the two traces initially on top of each other. With the latter initial condition, adjust the spacing until the darker area first seems to appear. Use the average of the two readings to compute your noise estimate, and also compute the difference between the two readings to provide an estimate of your measurement uncertainty. With care and a little practice, sub-1dB repeatability is readily achievable.

5. G. Franklin and T. Hatley, “Don’t Eyeball Noise,” *Electronic Design* 24, Nov. 22, 1973, pp. 184-187.

Narrowband LNA Design Lab

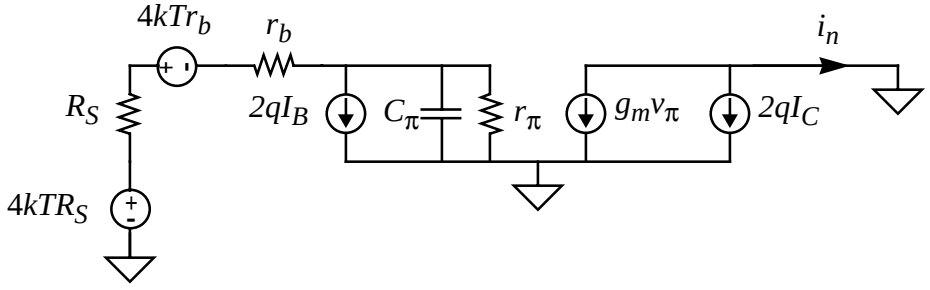
1.0 Introduction

In this lab experiment, you will design, construct and test a low noise amplifier for use at 1GHz. The design goals are: NF of ~2dB (if you can achieve 1.XdB, I want to see it), power gain of 10dB or more, and return loss greater than 10dB (both input and output). The 2SC3302 should be barely capable of the noise figure and power gain specified, so you can't be too terribly sloppy. The noise figure is competitive with what is achieved by many cell phones, by the way (although to be fair, they have more junk between the antenna and the amplifier). The return loss specifications are pretty generous, and you should be able to do substantially better than 10dB without much trouble.

2.0 Summary of Bipolar Noise Model

Be sure to read the handout on LNA design. Excerpted figures and equations are provided here to cut down on the amount of paper you have to haul around:

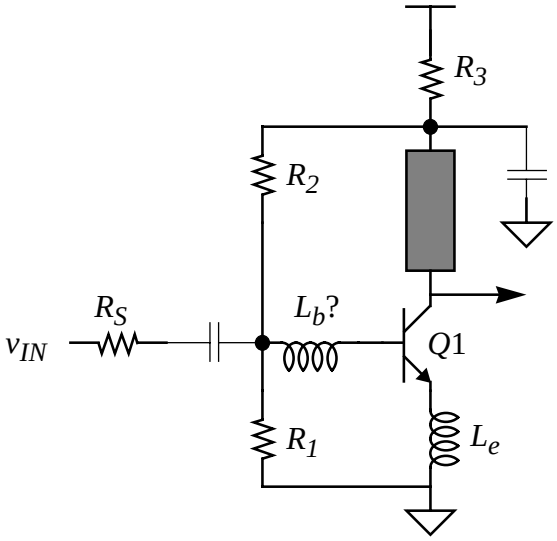
FIGURE 1. Model for noise figure calculation



$$F = 1 + \frac{r_b}{R_S} + \frac{(R_S + r_b)^2 + |z_\pi|^2 + 2(R_S + r_b)\,\mathrm{Re}\{z_\pi\}}{2R_S|z_\pi|^2 g_m} + \frac{(r_b + R_S)^2 g_m}{2\beta_F R_S}. \quad (1)$$

$$R_S^2 \approx \frac{\left(\frac{\omega_T}{\omega}\right)^2 \left(\frac{2r_b}{g_m} + \frac{r_b^2}{\beta_F} + \frac{1}{g_m^2}\right) + r_b^2}{1 + \frac{\left(\frac{\omega_T}{\omega}\right)^2}{\beta_F}}. \quad (2)$$

FIGURE 2. Narrowband LNA biasing details



Notice that the figure shows no input impedance transformer. In general, you will need one, so don't forget to design it in. The input capacitor (not shown in the notes) is needed to prevent the output of the generator from fighting the bias of your amplifier.

Other hints: Power attenuation ahead of the amplifier adds directly to noise figure, dB for dB. Since the minimum NF for the transistor is already around 1.7dB, you can't afford to throw away very much before you exceed the 2dB specification. So mount your transistor reasonably close to the input BNC, because you will be picking up close to 0.1dB of loss for every inch of line leading up to your amplifier. Leave enough room for your matching network, certainly, but not a whole lot more.

The collector bypass capacitor may have to be a parallel combination of two capacitors. Put a small-value capacitor as close to the line as possible (and with as short a path to ground as possible) for best high-frequency bypassing. Then parallel that with a higher value capacitor (say, 10 to 100nF) to take care of any lower frequency junk. That second capacitor can have longer path lengths to the collector circuit and ground.

Finally, the noise of the bias resistors will have the least effect if you can connect their common point (the junction of R_1 and R_2) to the lower impedance side of your matching transformer network. Of course, this only works if your transformer has a DC path to the base! (And make sure it doesn't short the base to ground at DC!)

Narrowband LNA Design

1.0 Introduction

The first stage of a receiver is typically a low-noise amplifier (LNA), whose main function is to provide enough gain to overcome the noise of subsequent stages (typically a mixer). Aside from providing this gain while adding as little noise as possible, an LNA should accommodate large signals without distortion, and frequently must also present a specific impedance, such as 50Ω, to the input source. This last consideration is particularly important if a filter precedes the LNA, since the transfer characteristics of many filters (both passive and active) are quite sensitive to the quality of the termination.

We will see that one can obtain the minimum noise figure from a given device by using a particular magic source impedance whose value depends on the characteristics of the device. Unfortunately this source impedance generally differs, perhaps considerably, from that which maximizes power gain. Hence it is possible for poor gain and a bad input match to accompany a good noise figure. One aim of this chapter is to place this tradeoff on a quantitative basis to assure a satisfactory design without painful iteration.

We will focus on a single narrowband LNA architecture that it is capable of delivering near-minimum noise figures, along with an excellent impedance match and reasonable power gain. The narrowband nature of the amplifier is not necessarily a liability, as many applications require filtering anyway. The LNA we'll study thus exhibits a balance of many desirable characteristics.

2.0 Derivation of a Bipolar Noise Model

Before we can appreciate the attributes (and limitations) of the narrowband LNA topology, it's necessary first to derive an appropriate noise model for a bipolar transistor. To make the analysis tractable and facilitate the acquisition of design insight, we'll need to make a number of simplifying assumptions. These assumptions are not seriously erroneous as long as the device is operated at frequencies well below (say, at least a factor of five below) f_T. At still higher frequencies, rapid degradation of other device characteristics (such as gain) militates against the use of the device in the first place, and therefore obviates the need for analysis, accurate or otherwise.

Each of the two junctions in a bipolar transistor produces *shot noise*, modeled by a shunt current source whose mean-square spectral density is $2qI_{DC}$, where I_{DC} is the value of the bias current through the junction. The shot noise currents from the two junctions may be treated as uncorrelated for most practical purposes, so we will ignore correlations in all that follows. This neglect will allow us to add noise *powers* directly. That is, a funny (and very useful) kind of superposition is enabled by invoking statistical independence of noise sources.

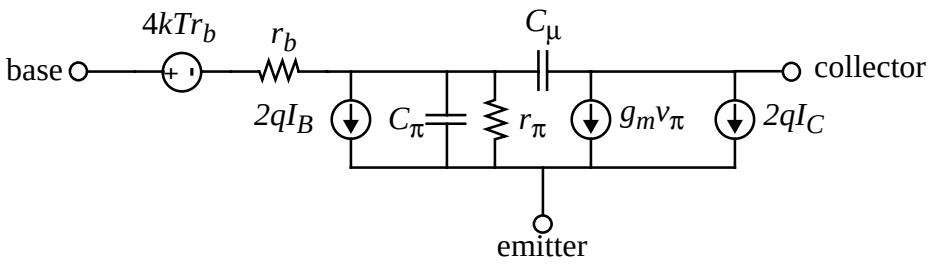
In addition to the shot noise components (which are in a sense fundamental, because no cleverness in device design can eliminate them), there is also a source of thermal noise: series base resistance, r_b. This noise is represented by a series voltage source whose mean-square density is $4ktr_b$. In modern devices its noise usually dominates (by a good margin) over that due to any series emitter or collector resistance, so we will neglect these. As we'll see, r_b is highly undesirable. Aside from generating noise (and thereby degrading noise figure), its presence often raises to inconvenient values the source resistance that yields minimum noise figure (as we'll see).

Although it is tempting to attribute thermal noise to all resistors appearing in a transistor model (e.g., r_π), doing so can amount to double counting. For example, r_π results from linearizing junction behavior, and junction noise is already modeled by shot noise. There is thus a difference between resistances that result from such linearization, and those that are simply ordinary resistors. The former do not generate thermal noise, while the latter do.

Finally, the collector-emitter output resistance is usually (but not always) large enough to be neglected at high frequencies, so we will omit it in all subsequent analyses.[1]

A small-signal transistor model based on these considerations appears as follows:

FIGURE 1. Noise model for bipolar transistor



This model, simple as it is, nonetheless captures the most important effects for calculating the noise figure of a bipolar amplifier. In fact we now have enough information to derive a usefully accurate expression for the noise figure of an amplifier, as well as to discover the value of the optimum source resistance.

Of the many possible ways to express noise factor, one that is especially useful here is:

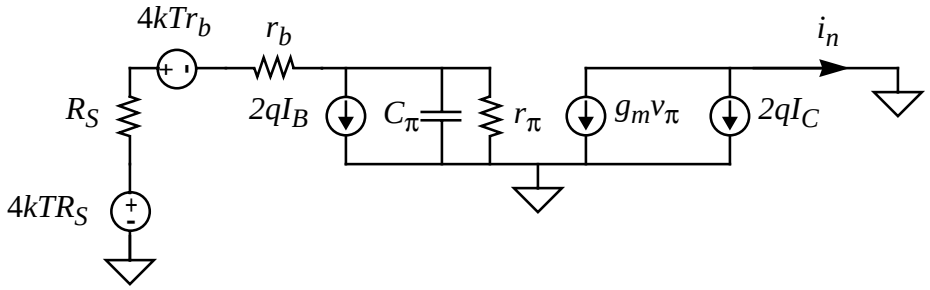
$$F \equiv \frac{\text{total output noise power}}{\text{output noise power due to source}}, \tag{1}$$

where, as usual, the source temperature is 290K.

1. The collector-emitter resistance models the Early effect, and is not thermally noisy because it is the result of linearizing the effect of junction-width variations. Finally, there is a collector-base feedback resistance that also arises from basewidth modulation. Its effect can almost always be completely ignored at RF.

To calculate the noise factor using Eqn. 1, connect a (thermally noisy) source resistance to the circuit of Fig. 1 and calculate away:

FIGURE 2. Model for noise figure calculation



Note that the circuit is terminated in a short. In an actual circuit, of course, the output would be loaded with a resistor of some nonzero value, unless the goal is to make a high-tech space heater. However it should be clear from Eqn. 1 that a collector load resistance appears as a multiplier in both the numerator and denominator. As a consequence, it ultimately cancels out and any value is okay for our purposes. We have therefore chosen a zero load resistance, a particularly convenient value.

A considerably sleazier trick is that we have arbitrarily eliminated the collector-base capacitance. Its presence complicates the analysis enough that its removal is necessary simply for clarity. As long as the collector load is a low impedance, this neglect is usually not too serious. In the general case, however, where arbitrary collector loads are to be considered, omitting C_μ can result in significant error. The largest error is in computing the source resistance that leads to the minimum noise figure. Depending on the detailed nature of the load impedance, the optimum source resistance could go up or down. Fortunately, the actual value of that minimum noise figure is usually not greatly affected, so we will proceed to derive the noise figure with a full awareness of the several assumptions that underlie its development.

Given those assumptions, and the use of a short-circuit load, then the noise factor is simply a ratio of short-circuit currents flowing in the collector branch labeled with i_n. The numerator is the sum of the mean-square short circuit currents due to all noise sources, and the denominator is the mean-square short circuit current due only to the source noise. Hence, we get

$$F = \frac{2qI_C + 2qI_B\left|z_\pi \| (r_b + R_S)\right|^2 g_m^2 + (4kTr_b + 4kTR_S)\left|\dfrac{z_\pi}{R_S + r_b + z_\pi}\right|^2 g_m^2}{4kTR_S\left|\dfrac{z_\pi}{R_S + r_b + z_\pi}\right|^2 g_m^2}, \qquad (2)$$

where the simple additions of terms in the numerator are a direct consequence of neglecting any correlations among the noise generators.

It's a good idea to study this equation one term at a time to try to make some sense of it. In the denominator, the mean-square voltage spectral density of the source resistor noise, $4kTR_S$, is first multiplied by the square of a voltage divide factor magnitude to find the mean-square voltage across r_π. That squared voltage in turn is multiplied by the square of the transconductance, g_m, to find the squared collector current and thus complete the denominator.

Examining the terms in the numerator from right to left, note that the noise voltage generator of resistor r_b is in series with that of R_S. It therefore undergoes precisely the same transformations, explaining why the last of the three additive terms in the numerator has the form shown.

The base shot noise current sees a total impedance that is a parallel combination of z_π (which, in turn, is r_π in parallel with C_π) and the sum $(r_b + R_S)$. Multiplying the mean-square shot noise current by the squared magnitude of that impedance gives us the mean-square voltage across r_π. Again multiplying that factor by the square of g_m yields the base shot noise contribution to the mean-square collector current.

Finally, the collector shot noise undergoes no scaling or other transformations at all, and so it adds directly to all of the other contributions in the numerator.

The equation can be simplified by cancelling some common terms, yielding (after a little re-ordering)

$$F = 1 + \frac{r_b}{R_S} + \frac{2qI_C + 2qI_B\left|z_\pi \| (r_b + R_S)\right|^2 g_m^2}{4kTR_S\left|\frac{z_\pi}{R_S + r_b + z_\pi}\right|^2 g_m^2}, \tag{3}$$

which simplifies still further to

$$F = 1 + \frac{r_b}{R_S} + \frac{2qI_C\left|R_S + r_b + z_\pi\right|^2}{4kTR_S\left|z_\pi\right|^2 g_m^2} + \frac{2qI_B\,(r_b + R_S)^2}{4kTR_S}. \tag{4}$$

We can continue to cancel common terms to obtain an even simpler form:

$$F = 1 + \frac{r_b}{R_S} + \frac{\left|R_S + r_b + z_\pi\right|^2}{2R_S\left|z_\pi\right|^2 g_m} + \frac{(r_b + R_S)^2 g_m}{2\beta_F R_S}. \tag{5}$$

In getting to this last expression, we have made use of the fact that the transconductance of a bipolar transistor is qI_C/kT, and that the ratio of collector to base current is β_F.

Note that the second term accounts for noise caused directly by the base resistance, the third term is due to collector shot noise, and the last term is the base current shot noise term. This is the last form of the equation that allows us to make these identifications.

Note also that Eqn. 5 contains three classes of terms (when everything is multiplied out). One is independent of R_S, another is proportional to R_S, and the third is inversely proportional to R_S. At very small source resistance the inversely proportional term dominates, and at very large values the proportional term dominates. Somewhere between "very small" and "very large" there is an optimum value that minimizes the sum (and, therefore, the noise figure). Before computing the optimum itself, let's understand intuitively why an optimum R_S should exist at all.

At very low source resistances, the contribution by the base resistance is more significant compared to that of the source itself, and noise figure therefore suffers.

At very high source resistances, the contribution to the output noise by the base shot noise is greater (because the impedance it faces is larger, generating a greater voltage across r_π, resulting in a greater current out of the collector). At the same time, the output noise due to the source itself is smaller, because of the harsher voltage divider seen by R_S. The magnitude of the collector shot noise does not change, but its size *relative* to the contribution by R_S is worse, so noise figure degrades further still. The optimum balances the contribution of the base resistance against the effects of base and collector shot noise.

The noise factor equation we will use is a slightly expanded version of Eqn. 5:

$$F = 1 + \frac{r_b}{R_S} + \frac{(R_S + r_b)^2 + |z_\pi|^2 + 2(R_S + r_b)\,\mathrm{Re}\{z_\pi\}}{2R_S|z_\pi|^2 g_m} + \frac{(r_b + R_S)^2 g_m}{2\beta_F R_S}. \tag{6}$$

Let's now do the math to derive the optimum value for R_S.

2.1 Optimum source resistance

The procedure for finding this optimum is straightforward enough: Take the first derivative with respect to the source resistance, set it equal to zero, and hope for a minimum:

$$\frac{d}{dR_S}\left(1 + \frac{r_b}{R_S} + \frac{|R_S + r_b + z_\pi|^2}{2R_S|z_\pi|^2 g_m} + \frac{(r_b + R_S)^2 g_m}{2\beta_F R_S}\right) = 0\ . \tag{7}$$

Grinding inexorably toward the answer generates the following sequence of equations:

$$\frac{d}{dR_S}\left(\frac{r_b}{R_S} + \frac{(R_S + r_b)^2 + |z_\pi|^2 + 2(R_S + r_b)\,\mathrm{Re}\{z_\pi\}}{2R_S|z_\pi|^2 g_m} + \frac{(r_b^2 + R_S^2 + 2r_b R_S)\,g_m}{2\beta_F R_S}\right) = 0\ ; \tag{8}$$

$$\frac{d}{dR_S}\left(\frac{r_b}{R_S} + \frac{R_S^2 + r_b^2 + |z_\pi|^2 + 2r_b\mathrm{Re}\{z_\pi\}}{2R_S|z_\pi|^2 g_m} + \frac{(r_b^2 + R_S^2)\,g_m}{2\beta_F R_S}\right) = 0\ , \tag{9}$$

where, in this last equation, we have taken out terms that are independent of R_S and therefore whose derivative is zero (we already took out the unity additive factor in getting to Eqn. 5, in case you were wondering where it went). (If you simply want to use the final answer, rather than follow each step of this derivation, feel free to skip ahead!)

Separating terms that are proportional to R_S from those that are inversely proportional to it leads us to the following:

$$\frac{d}{dR_S}\left[\frac{1}{R_S}\left(r_b+\frac{r_b^2+|z_\pi|^2+2r_b\mathrm{Re}\{z_\pi\}}{2|z_\pi|^2 g_m}+\frac{g_m r_b^2}{2\beta_F}\right)+R_S\left(\frac{1}{2|z_\pi|^2 g_m}+\frac{g_m}{2\beta_F}\right)\right]=0\ . \quad (10)$$

Taking the derivative at last, and setting it to zero yields

$$\left(\frac{1}{R_S^2}\right)\left(r_b+\frac{r_b^2+|z_\pi|^2+2r_b\mathrm{Re}\{z_\pi\}}{2|z_\pi|^2 g_m}+\frac{g_m r_b^2}{2\beta_F}\right)=\frac{1}{2|z_\pi|^2 g_m}+\frac{g_m}{2\beta_F}\ , \quad (11)$$

so that the optimum source resistance (squared) is

$$R_S^2=\frac{r_b+\dfrac{r_b^2+|z_\pi|^2+2r_b\mathrm{Re}\{z_\pi\}}{2|z_\pi|^2 g_m}+\dfrac{g_m r_b^2}{2\beta_F}}{\dfrac{1}{2|z_\pi|^2 g_m}+\dfrac{g_m}{2\beta_F}}, \quad (12)$$

which reduces a bit to

$$R_S^2=\frac{2|z_\pi|^2 g_m r_b+r_b^2+|z_\pi|^2+2r_b\mathrm{Re}\{z_\pi\}+\dfrac{g_m^2 r_b^2|z_\pi|^2}{\beta_F}}{1+\dfrac{|z_\pi|^2 g_m^2}{\beta_F}}. \quad (13)$$

This last equation is the last form that is traceable directly to our noise model without additional approximations. However, further simplification is possible if we allow one or two very reasonable approximations. One is that the operational frequency is well above $1/r_\pi C_\pi$ ($= \omega_T/\beta$), and the other is that the bias current is high enough that C_π is dominated by the diffusion capacitance. With these assumptions,

$$R_S^2 \approx \frac{\left(\frac{\omega_T}{\omega}\right)^2\left(\frac{2r_b}{g_m}+\frac{r_b^2}{\beta_F}+\frac{1}{g_m^2}\right)+r_b^2+\frac{2r_b r_\pi}{\left(\frac{\omega}{\omega_T}\right)^2\beta_F^2}}{1+\frac{\left(\frac{\omega_T}{\omega}\right)^2}{\beta_F}}. \tag{14}$$

If, as is often the case, the last term in the numerator is small compared to the term preceding it, we may write

$$R_S^2 \approx \frac{\left(\frac{\omega_T}{\omega}\right)^2\left(\frac{2r_b}{g_m}+\frac{r_b^2}{\beta_F}+\frac{1}{g_m^2}\right)+r_b^2}{1+\frac{\left(\frac{\omega_T}{\omega}\right)^2}{\beta_F}}. \tag{15}$$

As a specific numerical example, consider using a 2SC3302 microwave transistor at 1GHz. Assume that the collector bias current is 10mA, at which the transconductance is 400mS, $\beta = 80$, and ω_T is 10πGr/s. The remaining unknown is the value of r_b, which might remain unknown because it is rarely given in data sheets (the 2SC3302 is no exception). Fortunately, however, a plot of input impedance over frequency *is* given, and shows a resonance at approximately 800MHz when the bias current is 20mA (at which we may estimate C_π to be about 23pF, using other data sheet information). This resonance is the result of package and lead inductance interacting with C_π. Under that resonant condition, the input resistance is about midway on the Smith chart between the 25Ω and 50Ω contours, so we'll estimate the total resistance as 37–38Ω. This resistance is the sum of r_b and a real term produced by the series emitter inductance associated with the packaging and leads. As is shown in the next section, this induced resistance has a value $\omega_T L_e$. The parasitic inductance is not easily estimated but one can calculate from the resonance that the total inductance is approximately 1.7nH. This value is also quite believable from the physical dimensions of the package. Assuming that this total inductance splits evenly between base and emitter (even if it doesn't) allows us to estimate that the contribution by the induced resistance to the total is approximately 30Ω. Because this value is so close to the total estimated input resistance, our uncertainty in r_b is large. However, we'll press on, and use a value of 7-8Ω for r_b.

Under these conditions, the optimum source resistance is ~35Ω (at which the noise figure is 2dB at 1GHz), a value close enough to 50Ω that only a modest NF penalty (of a bit greater than 0.1dB) is incurred in this case if one performs no impedance transformation. At a bias current of 5mA, both the noise figure and the penalty for operating at 50Ω increase (the latter to about 0.2dB), for an overall minimum noise figure of about 3dB (again, this value is for operation at 1GHz).

As a check on our derivations, compare the calculated noise figure of 2dB to the minimum value of 1.7dB given in the data sheet for 1GHz operation. Repeating our calculation for 500MHz operation yields a 1.6dB noise figure (at 5mA), compared with a data sheet value of 1.5dB for that condition. Considering the crude nature of the approximations and parameter extractions, the overall level of agreement is satisfactory.

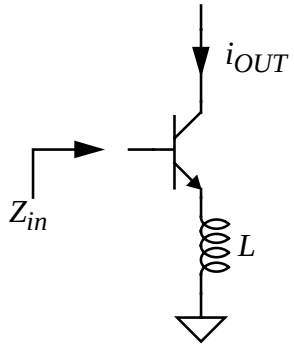
Finally, remember to keep in mind that an amplifier has to amplify. Achieving a low noise figure is important, but it is only half the battle. For this reason, selection of a suitable bias must take into account gain as well as noise figure. In the specific case of the 2SC3302, somewhat higher gain is obtained at the larger bias current, mainly because f_T is near its maximum value there. Since the minimum achievable noise figure does not change dramatically over the bias current range considered, there is considerable freedom, so other factors (such as implementation issues) may be taken into account as well.

3.0 The Narrowband LNA

The derivations of the previous section show that the source impedance that yields minimum noise factor is generally unrelated to the conditions that maximize power transfer. Furthermore, the input impedance of a bipolar transistor is intrinsically capacitive, so providing a good match to a 50Ω real source without degrading noise performance would appear difficult. Since presenting a known resistive impedance to the external world is almost always a critical requirement of LNAs, we will impose this requirement on our design as well.

A particularly good method for producing a real input impedance without degrading noise is to employ inductive emitter degeneration. With such an inductance, base current undergoes an additional phase shift beyond the ordinary quadrature relationship expected of a capacitor, causing the appearance of a resistive term in the input impedance. An important advantage of this method is that one has control over the value of the real part of the impedance through choice of inductance, as is clear from computing the input resistance of the following circuit:

FIGURE 3. Inductively-degenerated common-emitter amplifier



To simplify the analysis, consider a device model that includes only a transconductance and a base-emitter capacitance. In that case, it is not hard to show that the input impedance has the following form:

$$Z_{in} = sL + \frac{1}{sC_\pi} + \frac{g_m}{C_\pi}L \approx sL + \frac{1}{sC_\pi} + \omega_T L \quad . \tag{16}$$

Hence, the input impedance is that of a series *RLC* network, with a resistive term that is directly proportional to the inductance value.

More generally, an arbitrary source degeneration impedance *Z* is modified by a factor equal to [$\beta(j\omega)$ + 1] when reflected to the gate circuit, where $\beta(j\omega)$ is the current gain:

$$\beta(j\omega) = \frac{\omega_T}{j\omega}. \tag{17}$$

The current gain magnitude goes to unity at ω_T as it should, and has a capacitive phase angle because of C_π. Hence, for the general case,

$$Z_{in}(j\omega) = \frac{1}{j\omega C_\pi} + [\beta(j\omega) + 1]Z = \frac{1}{j\omega C_\pi} + Z + \left[\frac{\omega_T}{j\omega}\right]Z \ . \tag{18}$$

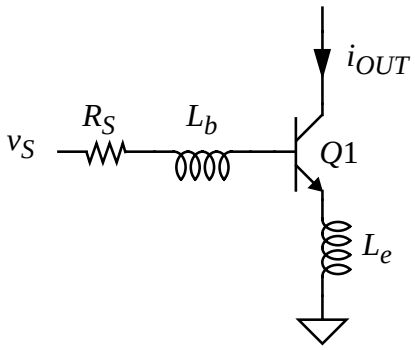
Note that capacitive degeneration contributes a *negative* resistance to the input impedance.[2] Hence, any parasitic capacitance from emitter to ground offsets the positive resistance from inductive degeneration. It is important to take this effect into account in any actual design (or use it to your advantage).

Whatever the value of this resistive term, it is important to emphasize that it does not bring with it the thermal noise of an ordinary resistor because a pure reactance is noiseless. We may therefore exploit this property to provide a specified input impedance without degrading the noise performance of the amplifier.

The form of Eqn. 16 clearly shows that the input impedance is purely resistive at only one frequency (at resonance), however, so this method can only provide a narrowband impedance match. Fortunately, there are numerous instances when narrowband operation is not only acceptable, but desirable, so inductive degeneration is certainly a valuable technique. The LNA topology we will examine for the rest of this chapter is therefore as follows:

2. Capacitively-loaded source followers are infamous for their poor stability. This negative input resistance is fundamentally responsible, and explains why adding some positive resistance in series with the gate circuit helps solve the problem.

FIGURE 4. Narrowband LNA with inductive emitter degeneration (biasing not shown)



The inductance L_e is chosen to provide the desired input resistance (equal to R_s, the source resistance). Since the input impedance is purely resistive only at resonance, an additional degree of freedom, provided by inductance L_b, is needed to guarantee this condition.[3] Now, at resonance, the base-to-emitter voltage is Q times as large as the input voltage. The overall stage transconductance G_m under this condition is therefore:

$$G_m = g_{m1}Q_{in} = \frac{g_{m1}}{\omega_o C_\pi (R_s + \omega_T L_e)} = \frac{\omega_T}{2\omega_o R_s}, \tag{19}$$

where we have used the approximation that ω_T is the ratio of g_{m1} to C_π.

The design procedure is thus reasonably straightforward. First select a bias current consistent with the gain and noise figure targets. Then compute the optimum source resistance to minimize noise figure. Next add enough emitter degeneration inductance to produce an input impedance whose real part is equal to the optimum source resistance, and then add enough of the right kind of impedance (e.g., more inductance) in the base circuit to remove any residual reactive input component and thereby bring the input loop into resonance. Finally, interpose a lossless matching network (if necessary) between the actual source and the amplifier to transform from 50Ω (or other source value) to the optimum value of R_S. This matching network often can be merged with whatever inductance (for example) is needed to resonate the input loop.

This particular procedure is attractive because it balances all parameters of interest. An excellent match is guaranteed by the inductive degeneration, while providing nearly the lowest noise figure possible at the given bias conditions. The resonant condition at the input also assures good gain at the same time, since the effective stage transconductance is proportional to ω_T/ω.

3. It can be the case that package and other parasitic inductance provides more than this value. In those cases, a series *capacitance* may be needed to resonate the input loop at the desired frequency.

4.0 A Few Practical Details

4.1 Realizing the emitter degeneration inductance

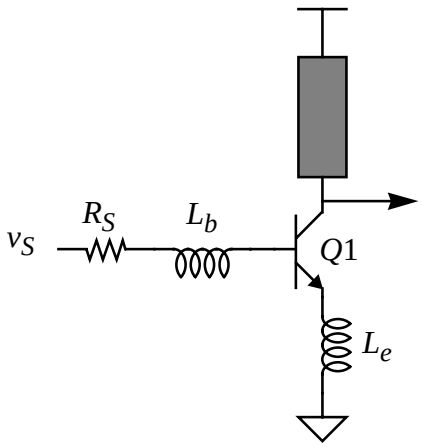
The narrowband LNA depends on inductive emitter degeneration to produce a real term in the input impedance. Quite often at microwave frequencies, the needed value is quite small, and therefore difficult to produce. For example, continuing with a 2SC3302 biased at 10mA, we would require ~2nH to produce 50Ω, and perhaps 2-3 times that inductance if the bias were reduced to 5mA (the increase in the impedance target for minimum noise figure, plus the reduction in ω_T, causes the needed inductance to increase faster than you might otherwise expect). However 2nH is not far from as small as one can expect to achieve without extreme measures, particularly since a fair fraction of this amount is already included in the packaging. Controlling the exact value is therefore challenging. In cases where the packaging and lead inductance already exceed the value you need, the input impedance will actually appear inductive, and thus require a capacitance to resonate the input loop. To avoid this necessity, extreme care in layout and construction is essential.

4.2 Collector load

It is generally the case that a resonant collector load is desired. Such a load increases gain by resonating out any output capacitance. Furthermore, the additional filtering of unwanted signals is highly desirable.

There are several practical options for realizing such a load. One is to use a discrete inductor of some appropriate value. A preferable one for our purposes is to implement the inductor out of a suitable length of microstrip, because of its versatility:

FIGURE 5. Narrowband LNA with microstrip load (biasing not shown)



The length is adjusted to produce resonance. And, if needed, a downward impedance transformation is readily obtained by merely tapping the output off of some intermediate position along the line. Clearly, the impedance is a minimum (zero) at the V_{CC} end of the

line, and a maximum at the collector end. To a first (and crude) approximation, the impedance varies quadratically along the line.

The sharpness of the resonance can be adjusted by varying the width of the line. The width controls the L/C ratio of the line, and therefore controls Q.

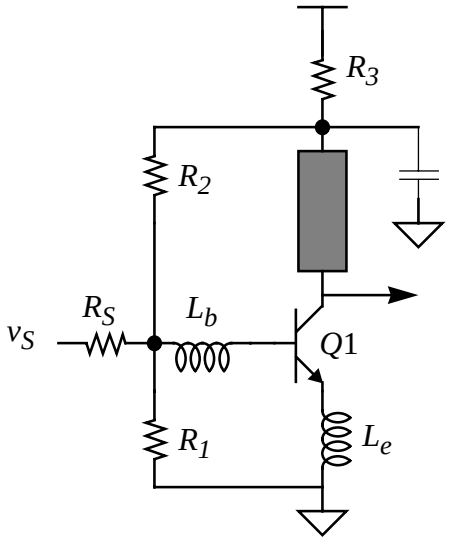
The other attractive attribute of the line is that it makes biasing relatively simple, as will be seen in the next section.

4.3 Biasing

There are numerous ways to bias a single-ended amplifier at low frequencies. Our options narrow somewhat at microwave frequencies because we cannot always tolerate the impedances that necessarily attend discrete implementations of bias networks. For example, it is very common at lower frequencies to bias the base through a voltage divider, and then insert a stabilizing emitter degeneration resistor. To buy back signal gain, a bypass capacitor is placed across this resistor.

In our case, it is probably not practical to use this approach because any “junk” in the emitter circuit only makes our job of implementing tiny inductances tougher. However, since the goal of resistive emitter degeneration is to reduce DC gain through negative feedback, we can seek alternative ways of accomplishing the same net goal. We can apply negative feedback to the base from the collector:

FIGURE 6. Narrowband LNA biasing details



The details of operation are left as an exercise for the reader, but a quick qualitative description is that the voltage across R_2 is a multiplied-up version of the voltage across R_1. The latter, in turn, is the base emitter voltage of the transistor. Since V_{BE} is temperature sensitive, so is the output voltage. However, the variation is small enough for our purposes that it is still a useful circuit.

The collector load resistor, R_3, is bypassed by a capacitor to keep the top of the microstrip load a reasonable signal ground. This bypassing need not be perfect, however, because additional inductance here only forces us to shorten the load a little bit. Moving the bias feedback takeoff point from the emitter to the collector thus solves a thorny problem.

As a final note on this bias method, the resistors have to be chosen small enough so that the current flowing through them is large compared with *variations* in transistor base current, if the bias point is to remain roughly insensitive to base current. This requirement is somewhat at odds with the desire to keep the resistors large to minimize their contribution to thermal noise. Fortunately, it is usually not difficult to find an acceptable compromise, and net degradations in noise figure can be kept to the level of tenths of a dB or less.

5.0 Linearity and Large-Signal Performance

In addition to noise figure, gain and input match, linearity is also an important consideration because an LNA must do more than simply amplify signals without adding much noise. It must also remain linear even when strong signals are being received. In particular, the LNA must maintain linear operation when receiving a weak signal in the presence of a strong interfering one, otherwise a variety of pathologies may result. These consequences of intermodulation distortion include desensitization (also known as blocking) and cross-modulation. Blocking occurs when the intermodulation products caused by the strong interferer swamp out the desired weak signal, while cross-modulation results when nonlinear interaction transfers the modulation of one signal to the carrier of another. Both effects are undesirable, of course, so another responsibility of the LNA designer is to mitigate these problems to the maximum practical extent.

The LNA design procedure described in this chapter does not address linearity directly, so we now develop some methods for evaluating the large-signal performance of amplifiers, with a focus on the acquisition of design insight. As we'll see, although the narrowband LNA topology achieves its good noise performance somewhat at the expense of linearity, the trade-off is not serious enough to prevent the realization of LNAs with more than enough dynamic range to satisfy demanding applications.

While there are many measures of linearity, the most commonly used are third-order intercept (IP3) and 1dB compression point (P_{1dB}).[4] To relate these measures to readily calculated circuit and device parameters, suppose that the amplifier's output signal may be represented by a power series.[5] Furthermore, assume that we will evaluate these measures with signals small enough that truncating the series after the cubic term introduces negligible error:

$$i(V_{DC} + v) \approx c_0 + c_1 v + c_2 v^2 + c_3 v^3, \tag{20}$$

4. In direct-conversion (homodyne) receivers, the second-order intercept is more important.

5. We are also assuming that input and output are related through an anhysteretic (memoryless) process. A more accurate method would employ Volterra series, for example, but the resulting complexity obscures much of the design insight we are seeking.

where Eqn. 20 describes the specific case of a transconductance.

Now consider two equal-amplitude sinusoidal input signals of slightly different frequencies:

$$v = A\,[\cos(\omega_1 t) + \cos(\omega_2 t)]\,. \tag{21}$$

Substituting Eqn. 21 into Eqn. 20 allows us, after simplification and collection of terms, to identify the components of the output spectrum.[6] The DC and fundamental components are as follows:

$$\left[c_0 + c_2 A^2\right] + \left[c_1 A + \frac{9}{4} c_3 A^3\right][\cos(\omega_1 t) + \cos(\omega_2 t)]\,. \tag{22}$$

Note that the quadratic factor in the expansion contributes a DC term that adds to the output bias. The cubic factor augments the fundamental term, but by a factor proportional to the cube of the amplitude, and thus contributes more than a simple increase in gain. In general, DC shifts come from even powers in the series expansion, while fundamental terms come from odd factors.

There are also second- and third-harmonic terms, caused by the quadratic and cubic factors in the series expansion, respectively:

$$\left[\frac{c_2 A^2}{2}\right][\cos(2\omega_1 t) + \cos(2\omega_2 t)] + \left[\frac{c_3 A^3}{4}\right][\cos(3\omega_1 t) + \cos(3\omega_2 t)]\,. \tag{23}$$

In general, *n*th harmonics come from *n*th-order factors. Harmonic distortion products, being of much higher frequencies than the fundamental, are usually attenuated enough in tuned amplifiers so that other nonlinear products dominate.

The quadratic term also contributes a second-order intermodulation (IM) product, as in a mixer (see next chapter):

$$\left[c_2 A^2\right][\cos(\omega_1 + \omega_2)\,t + \cos(\omega_1 - \omega_2)\,t]\,. \tag{24}$$

As with the harmonic distortion products, these sum and difference frequency terms are effectively attenuated in narrowband amplifiers if ω_1 and ω_2 are nearly equal, as assumed here.

Finally, the cubic term gives rise to third-order intermodulation products:

$$\left(\frac{3}{4} c_3 A^3\right)[\cos(\omega_1 + 2\omega_2)\,t + \cos(\omega_1 - 2\omega_2)\,t + \cos(2\omega_1 + \omega_2)\,t + \cos(2\omega_1 - \omega_2)\,t]\,. \tag{25}$$

6. This derivation makes considerable use of the following trigonometric identity:
(cosx)(cosy) = [cos(x+y) + cos(x-y)]/2.

Note that these products grow as the cube of the drive amplitude. In general, the amplitude of an *n*th-order IM product is proportional to the *n*th-power of the drive amplitude.

The sum frequency third-order IM terms are of diminished importance in tuned amplifiers because they typically lie far enough out of band to be significantly attenuated. The difference frequency components, however, can be quite troublesome since their frequencies may lie in band if ω_1 and ω_2 differ by only a small amount (as would be the case of a signal and an adjacent-channel interferer, for example). It is for this reason that the third-order intercept is an important measure of linearity.

It is straightforward from the foregoing sequence of equations to compute the input-referred third-order intercept (IIP3) by setting the amplitude of the IM3 products equal to the amplitude of the fundamental:

$$\left|c_1 A\right| = \left|\frac{3}{4}c_3 A^3\right| \rightarrow A^2 = \frac{4}{3}\left|\frac{c_1}{c_3}\right|, \tag{26}$$

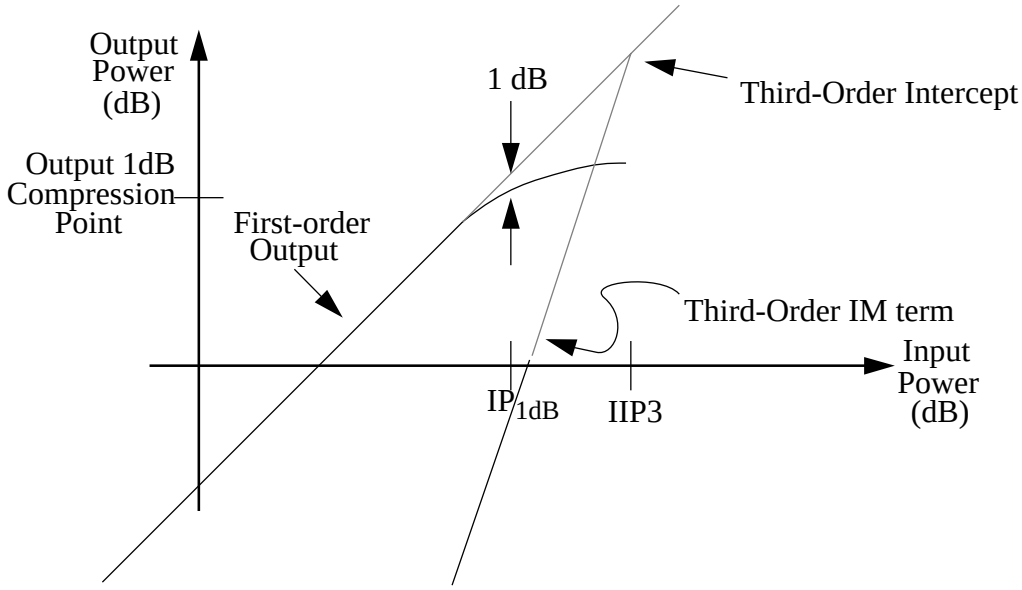
where we have assumed only a weak departure from linearity in expressing the fundamental output amplitude. It is important to emphasize that the intercept is an extrapolated value because the corresponding amplitudes computed from Eqn. 26 are almost always so large that truncating the series after the third-order term introduces significant error. In both simulations and experiment, the intercept is evaluated by extrapolating trends observed with relatively small amplitude inputs.

Since Eqn. 26 yields the square of the voltage amplitude, dividing by twice the input resistance R_s gives us the power at which the extrapolated equality of IM3 and fundamental terms occurs:

$$IIP3 = \frac{2}{3}\left|\frac{c_1}{c_3}\right|\frac{1}{R_s}. \tag{27}$$

The following figure summarizes the linearity definitions:

FIGURE 7. Illustration of LNA performance parameters



In this figure, it is customary to plot the output powers as a function of the power of each of the two (equal amplitude) input tones, rather than their sum.

Since third-order products grow as the cube of the drive amplitude, they have a slope that is three times that of the first-order output when plotted on logarithmic scales, as in the figure. Note that, in the figure, the 1dB compression point occurs at a lower input power than IIP3. This general relationship is nearly always the case (by a healthy margin) in practical amplifiers.

Having defined the linearity measures, we now consider ways to estimate IIP3, with and without the aid of Eqn. 27.

5.1 Methods for Estimating IP3

One way to find IP3 is through a transient simulation in which two sinusoidal input signals of equal amplitude and nearly equal frequency drive the amplifier. The third-order intermodulation products of the output spectrum are compared with the fundamental term as the input amplitude varies and the intercept computed.

While simple in principle, there are several significant practical difficulties with the method. First, since the distortion products may be several orders of magnitude smaller than the fundamental terms, numerical noise of the simulator can easily dominate the output unless exceptionally tight tolerances are imposed.[7] A closely related consideration is that the time steps must be small enough and *equally spaced* not to introduce artifacts in the output spectrum.[8] When these conditions are satisfied, the simulations typically execute quite slowly, and generate large output files.

7. The tolerances must be *much* tighter, in fact, than the “accurate” default options commonly offered.

Pure frequency-domain simulators (e.g., harmonic balance tools) can compute IP3 in much less time, but are currently less widely available than time-domain simulators such as SPICE.

Eqn. 27 offers a simple expression for the third-order intercept in terms of the ratio of two of the power series coefficients, and thus suggests an alternative method that might be suitable for hand calculations. While one is rarely given these coefficients directly, it is a straightforward matter to determine them if an analytical expression for the transfer characteristic is available. Even without such an expression, there is an extremely simple procedure, easily implemented in "ordinary" simulators such as SPICE, that allows rapid estimation of IP3. This technique, which we'll call the three-point method, exploits the fact that knowing the incremental gain at three different input amplitudes is sufficient to determine the three coefficients c_1, c_2 and c_3.[9]

To derive the three-point method, start with the series expansion that relates input and output:

$$i(V_{DC}+v) \approx c_0 + c_1 v + c_2 v^2 + c_3 v^3. \tag{28}$$

The incremental gain (transconductance) is the derivative of Eqn. 20:

$$g(v) \approx c_1 + 2c_2 v + 3c_3 v^2. \tag{29}$$

While any three different values of v would suffice in principle, particularly convenient ones are 0, V and -V, where these voltages are interpreted as deviations from the DC bias value. With those choices, one obtains the following expressions for the corresponding incremental gains:

$$g(0) \approx c_1, \tag{30}$$

$$g(V) \approx c_1 + 2c_2 V + 3c_3 V^2, \text{ and} \tag{31}$$

$$g(-V) \approx c_1 - 2c_2 V + 3c_3 V^2. \tag{32}$$

Solving for the coefficients yields

$$c_1 = g(0), \tag{33}$$

$$c_2 = \frac{g(V) - g(-V)}{4V}, \text{ and} \tag{34}$$

8. This requirement stems from the assumption, made by all FFT algorithms used by practical simulators, that the time samples are uniformly spaced.

9. This method is an adaptation of a classic technique from the vacuum tube era which allows estimation of *harmonic* distortion.

$$c_3 = \frac{g(V) + g(-V) - 2g(0)}{6V^2}. \tag{35}$$

Substituting into Eqn. 27 these last three equations for the coefficients then gives us the desired expression for IIP3 in terms of the three incremental gains:[10]

$$IIP3 = \frac{4V^2}{R_s} \cdot \left| \frac{g(0)}{g(V) + g(-V) - 2g(0)} \right|. \tag{36}$$

Finding IIP3 with Eqn. 36 is much faster than through a transient simulation because determining the incremental gains involves such little computation for either a simulator or a human. The three-point method is thus particularly valuable for rapidly estimating IIP3 in the early stages of a design.

6.0 Spurious-Free Dynamic Range (SFDR)

So far, we have identified two general limits on allowable input signal amplitudes. The noise figure defines a lower bound, while distortion sets an upper bound. Loosely speaking, then, amplifiers can accommodate signals ranging from the noise floor to some linearity limit. Using a dynamic range measure helps designers avoid the pitfall of improving one parameter (e.g., noise figure) while inadvertently destroying another.

This idea has been put on a quantitative basis through a parameter known as the spurious-free dynamic range (SFDR). The term “spurious” means “undesired,” and is often shortened to “spur.”[11] In the context of LNAs, it usually refers to the third-order products, but may occasionally apply to other undesired output spectral components.

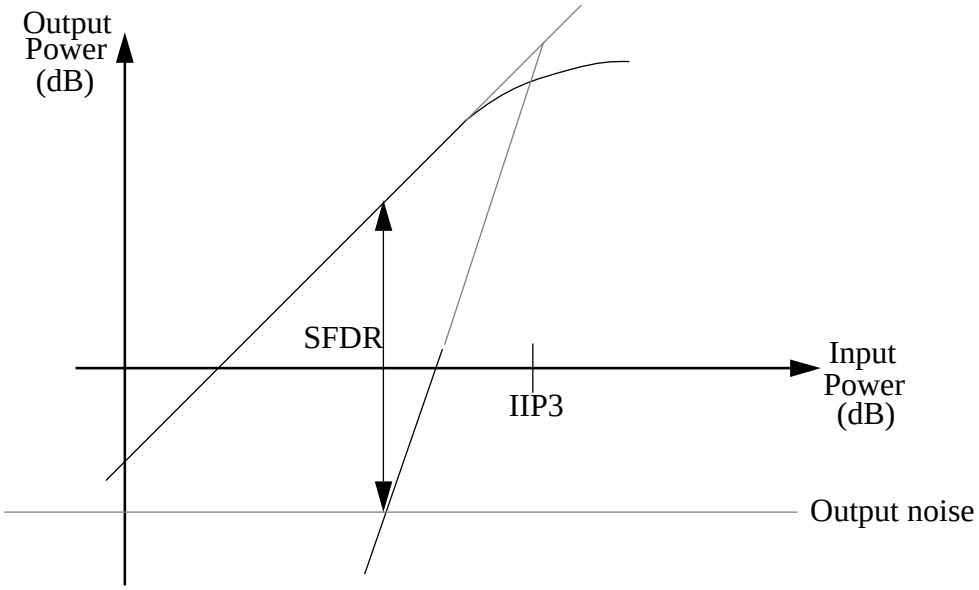
To understand the rationale behind using SFDR as a specific measure of dynamic range, define as a more general measure the lesser of signal-to-noise or signal-to-distortion ratio, and evaluate this measure as one varies the amplitude of the two tones applied to the amplifier. As the input amplitude increases from zero, the first-order output initially has a negative signal-to-noise ratio but eventually emerges from the noise floor. Because third-order distortion depends on the cube of the input amplitude, IM3 products will be well below the noise floor at this point for any practical amplifier. Hence, the dynamic range improves for a while as the input signal continues to increase, since the desired output increases while the undesired output (here, the noise) stays fixed. Eventually, however, the third-order IM terms also emerge from the noise floor. Beyond that input level, the dynamic range decreases, since the IM3 terms grow three times as fast (on a dB basis) as the first-order output.

10. Having determined all of the coefficients in terms of readily measured gains, it is easy to derive similar expressions for harmonic and second-order intermodulation distortion. The latter quantity is especially relevant for direct-conversion receivers.

11. Occasionally (and erroneously), “spurii” is used for the plural of “spurious,” even though “spurious” is not a Latin word. “Spurs” is the preferred plural.

The SFDR is defined as the signal-to-noise ratio corresponding to the input amplitude at which an undesired product (here, the third-order IM power) just equals the noise power, and is therefore the maximum dynamic range that an amplifier exhibits in the foregoing experiment, as is clear from the following figure:

FIGURE 8. Spurious-free dynamic range (third-order)



To incorporate explicitly the noise figure and IIP3 in an expression for SFDR, first define N_{oi} as the input-referred noise power in decibels. Then, since the third-order IM products have a slope of three on a dB scale, the input power below IIP3 at which the input-referred IM3 power equals N_o is given by (again, all powers are expressed in decibels):

$$\Delta P_1 = \frac{IIP3 - N_{oi}}{3}. \tag{37}$$

The SFDR is just the difference between the input power implied by Eqn. 37 and N_{oi}:

$$SFDR = (IIP3 - \Delta P_1) - N_{oi} = IIP3 - \frac{IIP3 - N_{oi}}{3} - N_{oi} = \frac{2}{3}(IIP3 - N_{oi}), \tag{38}$$

where, again, all power quantities are in decibels.

Note that the input-referred noise power (in watts this time) is simply the noise factor F times the noise power $kT\Delta f$. Note also that output-referred quantities may be used in Eqn. 38 because the same gain factor scales both terms.

It is satisfying that SFDR is indeed bounded on one end by IIP3, and on the other by the noise floor, as argued qualitatively at the beginning of this section. The factor 2/3 comes into play because of the particular way in which the limits are defined.

7.0 Summary

We've seen that an inductively degenerated LNA achieves simultaneously an excellent impedance match, nearly minimum noise figure, and reasonable gain.

The three-point method was also introduced, permitting an approximate, but quantitative, assessment of linearity in much less time than is possible with straightforward time-domain simulators. Even though the method neglects dynamics, measurements on practical amplifiers usually reveal reasonably good agreement with predictions. Reasonable agreement may generally be expected as long as the device is operated well below ω_T.

If better linearity is required, either power consumption or gain must degrade in exchange for the improved linearity. The bias conditions can be altered to decrease input Q, or negative feedback employed, for example. Finally, combining the signal amplitude limitations implied by the noise and distortion figures of merit yields a measure of the maximum dynamic range of an amplifier, the spurious-free dynamic range.

8.0 Appendix: Noise Figure Equations

Repeated here are the equations for optimum source resistance (both "exact" and approximate), and the corresponding noise factor:

$$R_S^2 = \frac{2\left|z_\pi\right|^2 g_m r_b + r_b^2 + \left|z_\pi\right|^2 + 2r_b \mathrm{Re}\left\{z_\pi\right\} + \dfrac{g_m^2 r_b^2 \left|z_\pi\right|^2}{\beta_F}}{1 + \dfrac{\left|z_\pi\right|^2 g_m^2}{\beta_F}}, \tag{39}$$

$$R_S^2 \approx \frac{\left(\dfrac{\omega_T}{\omega}\right)^2 \left(\dfrac{2r_b}{g_m} + \dfrac{r_b^2}{\beta_F} + \dfrac{1}{g_m^2}\right) + r_b^2 + \dfrac{2r_b r_\pi}{\left(\dfrac{\omega}{\omega_T}\right)^2 \beta_F^2}}{1 + \dfrac{\left(\dfrac{\omega_T}{\omega}\right)^2}{\beta_F}}, \text{ and} \tag{40}$$

$$F = 1 + \frac{r_b}{R_S} + \frac{\left(R_S + r_b\right)^2 + \left|z_\pi\right|^2 + 2\left(R_S + r_b\right)\mathrm{Re}\left\{z_\pi\right\}}{2R_S\left|z_\pi\right|^2 g_m} + \frac{\left(r_b + R_S\right)^2 g_m}{2\beta_F R_S}. \tag{41}$$

It might also be helpful to have the following expressions related to the impedance z_π:

$$\left|z_\pi\right| = \frac{r_\pi}{\sqrt{(\omega r_\pi C_\pi)^2 + 1}} \text{ and} \tag{42}$$

$$\text{Re}\{z_\pi\} = \frac{r_\pi}{(\omega r_\pi C_\pi)^2 + 1}. \tag{43}$$

Antennas

1.0 Introduction

It is important to remember that conventional lumped circuit theory results from approximating the way the universe behaves (in particular, from setting to zero some terms in Maxwell's equations, effectively making the speed of light infinite). The much vaunted "laws" of Kirchhoff[1] are not really laws at all; they are consequences of these approximations and, therefore, ultimately break down.[2] It turns out that the lumped descriptions of circuit theory, in which it is possible to identify elements as individual resistances, capacitances and inductances, are allowable only when the elements are small relative to a wavelength. Although a rigorous proof of this length criterion is somewhat outside of the spirit of a volume allegedly devoted to practical matters, perhaps a brief plausibility argument might suffice for the present.

If you are willing to accept as an axiom that the finiteness of the speed of light is not noticeable when the propagation delay T_D along a circuit element of length Δl is a small fraction of the shortest period of interest T_{min}, then we would require

$$\left(T_D = \frac{\Delta l}{v}\right) \ll T_{min} = \frac{1}{f_{max}}, \tag{1}$$

where v is the propagation velocity and f_{max} is the maximum frequency of interest. When rewritten, the inequality above may be expressed as

$$\Delta l \ll \frac{v}{f_{max}} = \lambda_{min}. \tag{2}$$

The wavelength λ_{min} is that of the highest frequency of interest.

Conventional circuit analysis is thus valid as long as circuit elements are "very small" compared to the shortest relevant wavelength. You might be tempted to argue that the restriction to "small" elements is not a serious practical constraint, because we may always subdivide a large structure into suitably small elements, each of which might be described accurately by a lumped approximation. However the problem with such an approach is that it quietly assumes that all the energy in the network is confined to the space occupied by the circuit elements themselves. In this chapter, we remove that assumption by allowing for the possibility of radiation of electromagnetic energy. In so doing, we identify the conditions which must be satisfied for significant radiation to occur. We shall see that radiation is theoretically possible from conductors of any length, but is

1. Please, two *h*'s and two *f*'s, and pronounced "keerk off" rather than "kirtch off."

2. Failure to acknowledge this fact is the source of an infinite variety of false conundrums, many of which are debated ad nauseam on various internet chat sites ("proof that physics is broken" and that sort of thing, written by folks who are often wrong, but never in doubt).

facilitated by structures whose dimensions are at least a significant fraction of a wavelength. Understanding this length dependency explains why we may almost always neglect radiation at low frequencies, and why practical antennas are as big as they are.

As with filters, the subject of antennas is much too vast for comprehensive treatment in just one chapter, of course.[3] The main goal here is to develop intuitive insights that are infrequently provided by (or perhaps difficult to extract from) rigorous mathematical treatments found in some texts, and to supplement the brief explanations commonly offered by many "how-to" books. Because this chapter is thus intended to complement, rather than replace, existing descriptions, be forewarned that we will sometimes (actually, often) sacrifice some rigor in favor of the development of design insight. In fact, there may be so much handwaving that you will occasionally need a windbreaker.

Aside from a refreshing breeze, the most important tangible product of such an approach is the development of simple analytical circuit models for antennas, and an appreciation of why there are so many different antenna configurations.

Although the book's focus is on planar circuits, we will begin with a study of the (electric) dipole antenna, not only because it is so widely used, but also because its analysis elucidates many issues common to all antennas. A clear understanding of a dipole's limitations explains why certain modifications, such as the addition of "capacity hats" or loading coils, can greatly improve the radiation properties of short dipoles. As will be shown, this same understanding reveals a relationship among normalized length, efficiency and achievable bandwidth that is reminiscent of the gain-bandwidth tradeoffs found in many amplifiers.

Equations describing the dipole also lead directly to a description of the magnetic loop antenna because they are duals; the loop antenna is a magnetic dipole antenna. In keeping with our planar viewpoint, the chapter spends a fair amount of time examining the microstrip patch antenna, which has become extremely popular in recent years because it is easily made with the same low-cost mass-production techniques that are used to make printed circuit boards. As will be seen, the intuitive foundations established during a study of the dipole serve well in understanding the patch antenna.

Because our preoccupation will be with equivalent circuit models for antennas, other important characteristics, such as radiation patterns, directivity and gain are sadly omitted here. The interested reader is directed to the references cited in Footnote 3 for excellent treatments of these topics.

3. Three excellent texts on this topic are *Antenna Theory and Design*, by Stutzman and Thiele (Wiley), the classic *Antennas*, by Kraus, and *Antenna Theory* (2nd ed.), by Constantine A. Balanis (Wiley). Much practical information on antenna construction for amateur radio work may be found in *The ARRL Antenna Handbook* and numerous other books by the ARRL (American Radio Relay League).

2.0 Poynting's Theorem, Energy and Wires

To develop a unified viewpoint that explains when a wire is a wire, and when it's an antenna, it is critically important to discard the mental imagery of electricity-as-a-fluid traveling down wires-as-pipes, that is consciously implanted in students before and during their undergraduate education. Instead understand that ideal wires, strictly speaking, *do not carry electromagnetic energy at all*. Many (perhaps most) students and engineers, to say nothing of lay people, find this statement somewhat controversial. Nevertheless, the statement that wires do not carry energy is correct, and it is easy to show.

To do so, start with the formula for power from ordinary low frequency circuit theory:

$$P = \frac{1}{2} Re\{ VI^* \}. \tag{3}$$

In simple words, delivery of real power requires voltage, current, *and* the right phase relationship between them (the asterisk in Eqn. 3 denotes complex conjugation). If either V or I is zero, no power can be delivered to a load. Furthermore, even if both are nonzero, a pure quadrature (90° phase) relationship still results in an inability to deliver real power.

The corresponding field theoretical expression of the same ideas is Poynting's theorem, which states that the (real) power associated with an electromagnetic wave is proportional to the vector cross product of the electric and magnetic fields:

$$P = \frac{1}{2} Re\{ \vec{E} \times \vec{H}^* \}. \tag{4}$$

To deliver real power, one must have E, H, and the right phase between them. If either E or H is zero, or if they are in precise quadrature, no power can be delivered. Now, the electric field inside a perfect conductor is zero. So, by Poynting's theorem, no (real) energy flows inside such a wire; if there is to be any energy flow, it must take place entirely in the space *outside* of the wire.[4] Many students who comfortably and correctly manipulate Poynting's theorem to solve advanced graduate problems in field theory nonetheless have a tough time when this particular necessary consequence is expressed in words, for it seems to defy common sense and ordinary experience ("I get a shock only when I *touch* the wire").

The resolution to this seeming paradox is that conductors *guide* the flow of electromagnetic energy. This answer may seem like semantic hair-splitting, but it is actually a profound insight that will help us to develop a unified understanding of wires, antennas, cables, waveguides, and even optical fibers. So for the balance of this text (and of your professional careers), retain this idea of conductors as guides, rather than conduits, for the electromagnetic energy that otherwise pervades space. Then many apparently different

4. This argument changes little when real conductors are considered. In that case, all that happens is the appearance of a small tangential component of electric field, which is just large enough to account for ohmic loss.

ways to deliver electromagnetic energy will be properly understood simply as variations on a single theme.

3.0 The Nature of Radiation

More than a few students have caught on to the fact that electrodynamic equations, rife with gradient, divergence and curl, are a devious invention calculated to torment hapless undergraduates. And from a professor's perspective, that is unquestionably the most valuable attribute of E&M (S&M?) theory.

But perhaps understandably, the cerebral hemorrhaging associated with this trauma frequently causes students to overlook important questions: What is radiation, exactly? How does a piece of wire know when and when not to behave as an antenna? What are the terminal electrical characteristics of an antenna? How are these affected by proximity to objects? Who invented liquid soap, and why?[5]

Let's begin with a familiar example from lumped circuit theory. Without loss of generality, consider driving a pure reactance (e.g. a lossless capacitor or inductor) with a sinusoidal source. If we examine the relationship between voltage and current, we find that they are precisely in time quadrature ("ELI the ICE man" and all that).[6] The *average* power delivered by the source to any pure reactance is zero because energy simply flows back and forth between the source and the reactance. In one quarter cycle, say, some amount of energy flows to the reactance, and in the next, that entire amount returns to the source. To deliver nonzero average power requires that there be an in-phase component of voltage and current. Adding a resistance across, or in series with, a reactance produces a shift in phase from a pure 90° relationship, producing just such an in-phase component and an associated power dissipation.

The question of power flow in electromagnetic fields involves precisely the same considerations. Whenever the electric and magnetic fields are in precise time quadrature, there can be no real power flow. If we define radiation as the conveyance of power to a remotely located load, lack of real power flow therefore implies a lack of radiation. We already know from lumped circuit theory that quadrature relationships prevail in nominal reactances at frequencies where all circuit dimensions are very short compared with the shortest wavelength of interest. For example, we treat as an inductance a short length of wire connected to ground, and as a capacitance a conductor suspended above a ground plane. These treatments are possible because the fields surrounding the conductors are changing "slowly."

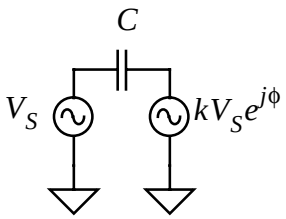
5. John Cusack in *The Sure Thing*, Embassy Films Associates, Rob Reiner, director, 1985.

6. Just in case this mnemonic is unfamiliar to you, it is a way of keeping track of the impedance phase relations in inductors and capacitors. "ELI" tells us that *E* comes before (leads) *I* in inductors, and "ICE" tells us that *I* leads *E* in capacitors.

Just as with those lumped reactance examples, real power delivery requires other than a pure 90° phase relationship between the electric and magnetic fields. To produce such a departure requires only the assistance of the finite speed of light to add extra delay.[7]

To understand concretely how the finite speed of light helps produce (actually, enables) radiation, consider a finite length of conductor driven at one end, say, by a sinusoidal voltage source. Near the source, the magnetic and electric fields may be well approximated as in quadrature. However, because it takes nonzero time for the signal to propagate along the conductor, the voltage (and, hence, the electric field) at the tip of the conductor is somewhat delayed relative to the voltage and electric field at the driven end. The currents (and their associated magnetic fields) at the two ends are similarly shifted in time. Thus the electric field at the far end is no longer precisely 90° out of phase with the magnetic field at the source end, and nonzero average power is consequently delivered by the driving source. A lumped circuit analogy that exhibits qualitatively similar features is the following network in which a capacitance is driven by two sinusoidal generators of equal frequency:

FIGURE 1. Capacitance driven by two isochronous sources



From a casual inspection of this particular network, one might be tempted to assert that there can be no power dissipation because a capacitor is a pure reactance. In fact, this is the most common answer given by prospective Ph.D. candidates during qualifying examinations. Let's directly evaluate the correctness of this assertion by computing the impedance seen by, say, the left source. The current through the capacitor is simply the voltage across it, divided by the capacitive impedance. So,

$$Z_{eq} = \frac{V_S}{V_S\,(1 - ke^{j\phi})\,sC} = \frac{1}{sC\,(1 - ke^{j\phi})} = \frac{1}{j\omega C\,[\,1 - k\,(\cos\phi + j\sin\phi)\,]}. \qquad (5)$$

The constant k is any real value (it is not meant to represent Boltzmann's constant). Notice that the factor in brackets has a purely real value whenever the phase angle ϕ is either zero or 180°. Under those conditions, the phase angle of the impedance is ±90°, implying zero dissipation. Energy is simply stored in the capacitor in one quarter cycle, then returned to the sources in the next. *Any* other phase angle produces the equivalent of a real component to the impedance as seen by the sources. Despite the presence of a pure reactance, dissipation is nonetheless possible. The capacitor certainly continues to dissipate zero power, but there are still two sources to consider. A nonzero average power transfer between these

7. Of course, this departure from quadrature also occurs when a real resistance is in the circuit. Our focus here is on radiation, so we will not consider dissipative mechanisms any further.

isochronous (i.e., equal frequency) sources is possible. That is, one source can perform work on the other.

Analogous ideas apply to the radiation problem. Because of the finiteness of propagation velocity, the electric and magnetic field components that normally simply store energy in the space around the conductors, suddenly become capable of delivering real power to some remotely located load; this is radiation. As a consequence, the signal source that drives the conductor must see the equivalent of a resistance, in addition to any reactance that might be present. One way to think about it is that this resistance, and radiation, *result from work performed by moving charges in one part of the antenna on charges in other parts of the antenna*.[8] The fields associated with radiation are actually present all the time (energy isn't in the conductors, it's in space), but radiation results only when the proper phase relationships exist.

From the foregoing description of radiation, it is also not difficult to understand why the length of an antenna is important. If the conductor (antenna) is very short, the time delay will be very short, leading to negligible departure from quadrature. More precisely, when the length of the conductor is very small *compared to a wavelength*, the resistive component of the antenna impedance will be correspondingly small. Normalization by the wavelength makes sense because a given length produces a fixed amount of time delay, and this time delay in turn represents a linearly increasing phase shift as frequency increases (wavelength decreases).

Now that we have deduced that radiation is a necessary consequence of a lack of pure quadrature, let us see if we can deduce the distance dependency of the radiation. Recalling that the electric field of an isolated, stationary charge in free space falls as the inverse square of distance, we might be tempted to argue that radiation must also exhibit an inverse square law.[9] To test this idea (again with a minimum of field theory), suppose we have a source of electromagnetic energy (it is completely unnecessary at this point to be more specific). Let's follow the outward flow of energy from the source through two successive (and concentric) spheres. If there is to be radiation, the total energy passing through the two spherical shells must be equal, or else the total energy would increase or decrease with distance, implying destruction or creation of energy.[10] We may therefore write:

$$\text{Energy} = \overline{P}_1 A_1 = \overline{P}_2 A_2, \tag{6}$$

where $\overline{P}$ is the areal power density, and A is the area. Now, because the surface area of a sphere is proportional to the square of the radius, constancy of total energy implies that the

8. Richard Feynman, the late Caltech physics Nobelist, described the process most succinctly of all: "If you shake an electron, light comes out." That is, radiation not only results from the fields of accelerated charges acting on the fields of other charges (either in the antenna or in surrounding media), but also may result from the action of an accelerated charge acting on its own field.

9. Because there's no such thing as absolute velocity, we may anticipate from elementary relativity considerations that radiation cannot result from a uniform motion of charge; acceleration is required.

10. Or a monotonically increasing storage of energy in free space, which we also disallow.

power density must decrease with the inverse distance squared. In free space, the electric and magnetic fields are proportional to each other. Coupling this fact with Poynting's theorem, we know that the power density is proportional to the square of the field strength,

$$\bar{P} \propto |E|^2 \propto |H|^2 \propto \frac{1}{r^2}. \tag{7}$$

Hence, we see that there must exist a component of electric or magnetic field whose amplitude falls as the *first* power of distance in order for radiation to be possible.[11] This development is remarkable, for if we had to depend solely on fields with an inverse-square spatial dependence (such as that of an isolated stationary charge), long-distance communications would be very difficult indeed (a $1/r^4$ power rolloff would be a catastrophe). Fortunately, a miracle of electrodynamics produces components of time-varying electric and magnetic fields that roll off much less dramatically (again, in free space). These radiation components are what make wireless communications practical. Although we certainly have not derived the precise form of the fields, we have nonetheless deduced important facts about them from very elementary arguments.

In addition to allowing us to associate radiation with the existence of inverse-distance fields, the foregoing tells us that the radiation of energy must be indistinguishable from energy dissipated in a resistor, from the point of view of the source. Correspondingly, we shall see that radiation contributes a resistive component to an antenna's driving point impedance, as asserted earlier.

Note also that the foregoing development actually makes a rather strong statement: *no* distance dependency other than inverse-distance can be associated with free space radiation. For example, if the fields were to fall off more rapidly, energy would have to accumulate in the space between two successive concentric spheres. If the fields decayed more slowly, energy would have to be supplied by that space. Since neither of these two conditions is compatible with the steady state, we conclude that such field components cannot support radiation. Instead, those other components must represent, at best, stored (reactive) energy, which flows back and forth between the source and the surrounding volume. Thus, their effect is accounted for with either inductive or capacitive elements in an antenna's circuit model, the development of which we will turn to shortly.

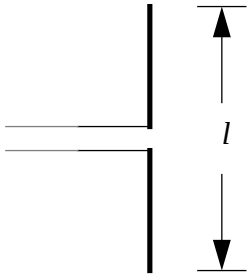
Having extracted about as much insight as is possible without resorting to any higher mathematics, we now turn to the practical problem of constructing and modeling real antennas.

4.0 The Dipole Antenna

The most common antenna is without question the dipole:

11. It is important to keep in mind that this conclusion depends on the assumption of free space propagation. If this assumption is violated (for example, by the presence of lossy media), then other conclusions may apply.

FIGURE 2. Short center-fed dipole antenna



Countless millions of dipoles in the form of “rabbit ears” have sat on top of television sets for decades, and countless millions more are presently found in cell phones and on automobiles. As we’ll see, the dipole operates on principles that follow directly the description of radiation we’ve given.

4.1 Radiation resistance

As one might suspect, the resistive equivalent of radiation is an extremely important parameter for an antenna. This radiation resistance determines, for example, how effectively energy from a source can be coupled into radiated energy. At the same time, by the principle of reciprocity, an antenna’s circuit model as a transmitter is the same as when the antenna is used as a receiver.

To derive the radiation resistance of a dipole from first principles is difficult enough that such antennas were used for a long time before such a derivation was actually carried out. An extremely useful engineering approximation is readily derived, however, by simply *assuming* a current distribution along the antenna. Theorists were guided toward a reasonable assumption by thinking about the dipole as approximately a two-wire transmission line (okay, it’s a bent one, but why be so picky?) that is terminated in an open circuit. Then an approximately sinusoidal current distribution results, with a boundary condition of nearly zero current at the open end of the wire.[12]

Using this assumed current distribution, and the assumption that the antenna is made of infinitesimally thin superconductors, one can derive the following approximation of the radiation resistance of a short dipole:

$$R_r \approx 20\pi^2 \left(\frac{l}{\lambda}\right)^2 . \tag{8}$$

The formula provides reasonably accurate answers for l/λ up to about 0.3. At a half-wavelength, the formula predicts a radiation resistance of about 50Ω, compared to an actual

12. The current doesn’t quite go to zero at the end because there is some nonzero fringing capacitance, but assuming that it does go to zero incurs a small enough error that the subsequent derivation is usefully accurate.

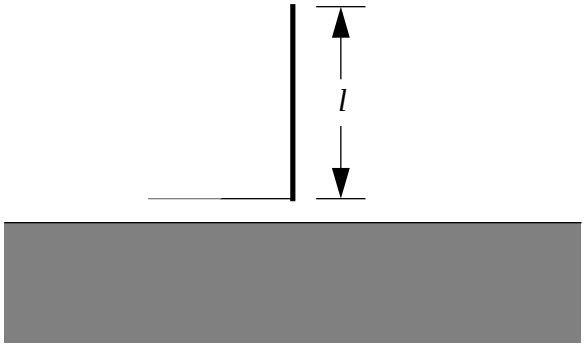
value of about 73Ω.[13] This value is a good match for the 75Ω world of video equipment, but some sort of matching network is needed for use in 50Ω systems.

The accuracy of the formula can be improved a little bit if one treats the antenna as slightly longer than its physical length. This effective extension results from the fact that the current along the antenna doesn't quite go to zero at the tip because of fringing field. We will later purposefully enhance this effect to improve (increase) the radiation resistance of short dipoles. In any event, a more accurate set of formulas is given in Table 1.

The length correction factor is a somewhat complicated, but rather weak, function of the radius-to-wavelength ratio, and is commonly taken as approximately 5% for typical dipole antennas. That is, the physical length should be multiplied by roughly 1.05, and that product inserted into Eqn. 8.

One of Marconi's key inventions is a valuable variation on the dipole antenna, in which image charges induced in the earth (or other conducting plane) effectively double the length of the antenna. For such a vertical monopole antenna over an ever-elusive perfect ground plane, the radiation resistance will be precisely double the value given by Eqn. 8.[14]

FIGURE 3. Short monopole antenna over ground plane



The doubling of effective length contributes a quadrupling of the radiation resistance. However, only the real vertical monopole (not the image) actually radiates, halving that quadrupling (got that?). The radiation resistance of a short monopole antenna is thus:

$$R_r \approx 40\pi^2 \left(\frac{l}{\lambda}\right)^2 . \tag{9}$$

This equation is reasonably accurate for l/λ values up to about 0.15-0.2. A quarter wavelength monopole will have an impedance of approximately 37Ω, compared to the formula's prediction of 25Ω. The length correction is again ~5%.

13. With conductors of finite diameter, the impedance is typically about 5% lower than this value.

14. A monopole is often also called a dipole antenna, because their operating principles and current distribution are fundamentally the same. In this text, we will use both terms, with the precise meaning to be inferred from context.

As does an open-circuited transmission line, both the monopole and dipole exhibit periodic resonances. The functional form of the radiation resistance varies somewhat as a function of resonant mode. For the center-fed dipole, approximate equations for the radiation resistance are presented in the following table:[15]

TABLE 1. Approximate radiation resistance for short and medium-length center-fed dipoles

Normalized conductor length, l/λ	R_{rad}
0 – 0.25	$20\left(\pi\frac{l}{\lambda}\right)^2$
0.25 – 0.5	$24.7\left(\pi\frac{l}{\lambda}\right)^{2.4}$
0.5 – 0.64	$11.1\left(\pi\frac{l}{\lambda}\right)^{4.17}$

The formulas of this table apply equally well to monopoles by doubling the constant multiplicative factor, and for normalized lengths that are half the values given in the first column. Again, the length is that of the actual conductor, not including the image.

4.2 Reactive components of antenna impedance

As noted earlier, radiation carries energy away, so its effect is modeled with a resistance. In general, however, some energy also generally remains in the vicinity of the antenna, flowing back and forth between the source and the surrounding volume. This near-field non-radiative component represents stored energy, and therefore contributes an imaginary component to the terminal impedance.

To derive highly approximate expressions for the effective reactance (inductance and capacitance) of short antennas, we again use the idea that a dipole antenna behaves much like an open-circuited transmission line. If we assume TEM propagation and unit values of relative permittivity and permeability, then the speed of light is expressed as

$$c = \frac{1}{\sqrt{LC}}, \tag{10}$$

where L and C here are the inductance and capacitance *per length*. Now, we already have the following equation for the approximate inductance per length of a wire with circular cross-section (see the chapter on passive components):

15. Stutzman and Thiele, *op. cit*, page 171.

$$L \approx \frac{\mu_0}{2\pi}\left[ln\left(\frac{2l}{r}\right) - 0.75 \right]. \tag{11}$$

The capacitance per unit length (in farads per meter) is thus very approximately:

$$C \approx \frac{1}{c^2 \frac{\mu_0}{2\pi}\left[ln\left(\frac{2l}{r}\right) - 0.75 \right]} \approx \frac{2\pi\varepsilon_0}{ln\left(\frac{2l}{r}\right) - 0.75} \approx \frac{5.56\times10^{-11}}{ln\left(\frac{2l}{r}\right) - 0.75}. \tag{12}$$

For typical dimensions, the capacitance per length is *very* roughly of the order of 10pF/m. For example, a 10cm length of 18 gauge conductor (about 1mm in radius) has a capacitance of almost exactly 1pF, according to the formula. Note that the inductance grows somewhat faster, and the capacitance somewhat more slowly, than linearly with length. Thus the capacitance per length is not a constant, but the 10pF/m estimate serves well for back-of-the-envelope calculations.

Again treating the dipole antenna as an open-circuited transmission line, we expect short dipoles to exhibit a primarily capacitive reactance, changing to a pure resistance as we lengthen the line toward resonance (at half wavelength), then to an inductance as we pass resonance. This general trend is periodic, repeating every wavelength, but the peak-to-peak variation in impedance diminishes because of the increasing loss.

The foregoing equations apply to the case of a short center-fed dipole. For a short end-fed monopole over a ground plane, the capacitance will be precisely double this amount. This result may be understood by recognizing that there exists a plane of symmetry between the two dipole segments. Interposing a grounded conducting plane at this location thus changes nothing. The capacitance of a center-fed dipole may thus be considered the result of two capacitances in series. Since each of these connects to ground, each series capacitance is in fact the value of the monopole capacitance. A similar argument allows us to deduce that the inductance of a monopole is one-half that for the center-fed dipole. Finally, we may also infer that the driving-point impedance of a monopole changes periodically every *half* wavelength.[16]

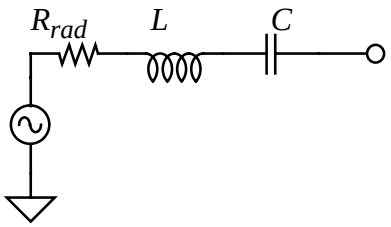
As a final comment, it must be reiterated that all of these equations assume that the antenna is in free space, without any other objects nearby (except for a ground plane in the case of the monopole). Measurements on real antennas often show significant deviations from the predictions of these simple equations, partly because of the simplemindedness underlying their derivation, but mainly because one is rarely able in practical circumstances to arrange for all objects to be very far removed from the antenna. Objects less than a few wavelengths away from the antenna can have an important influence on both the reactance and the radiation resistance. Loosely and unreliably speaking, antenna reactance is primarily sensitive to the proximity of dielectric substances (if the antenna is pri-

16. Again, loss due to radiation and any other dissipative mechanism causes the variation in impedance to diminish.

marily dependent on electric field) or of magnetic substances (if the antenna is primarily dependent on the magnetic field). The real term is generally most sensitive to nearby lossy substances.

Summarizing the results of this section, simple lossless dipoles may be modeled by the following simple circuit:

FIGURE 4. Circuit model for dipole antenna (one mode only)



In this model the generator represents the voltage induced by a received signal. When the antenna is used as a transmitter, the generator is set to zero value (a short). Any loss (arising, say, from skin effect), would be modeled by an additional resistance in series with the radiation resistance.

4.3 Capacitively loaded dipole

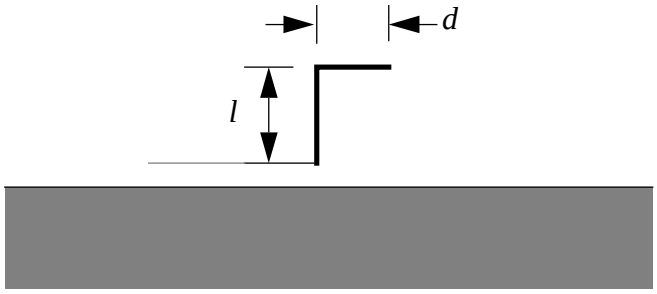
We've seen that the radiation resistance of a short dipole varies as the square of normalized length. Hence, good radiation requires a dipole to be a reasonable fraction of a wavelength, or else the radiation resistance will be too low to permit coupling energy into (or out of) it with high efficiency. Unfortunately, it is not always practical to lengthen an antenna arbitrarily to satisfy this requirement, particularly at low frequencies (remember, the free space wavelength at 1MHz is 300 meters). Sometimes an important constraint is imposed by a mechanical engineering problem, that of supporting a tall, skinny thing.

One way to finesse the problem is to bend the antenna (it's easier to support a long horizontal thing than a tall vertical thing). To understand why this is potentially beneficial, recall the observation that the fringing field of a straight dipole causes the antenna to act somewhat longer than its physical length. The capacitance associated with the fringing field prevents the current from going all the way to zero at the end, increasing the average current along the antenna, thus raising the radiation resistance. Although the effect is normally small, resulting in a length correction of only ~5% for ordinary dipole antennas, fringing can be purposefully enhanced to make short dipoles act significantly longer.[17] In applications where longer dipoles are not permitted because of space limitations in the vertical dimension, one can employ capacitive loading, using what are known as *capacity (or capacitive) hats*, to increase both the current at the end as well as the average current over the length of the dipole. Various conductor arrangements may be, and have been,

17. It may be shown that the absolute theoretical maximum impedance boost factor is four, corresponding to a constant amplitude current all along the dipole. In practice, the boost factors achieved are considerably smaller than allowed by theory.

used, including flat disks, spherical balls, and horizontal wires (the latter is used in the L- and T-antenna). Alas, accurate equations for these different cases are not easily derived.

FIGURE 5. L-antenna



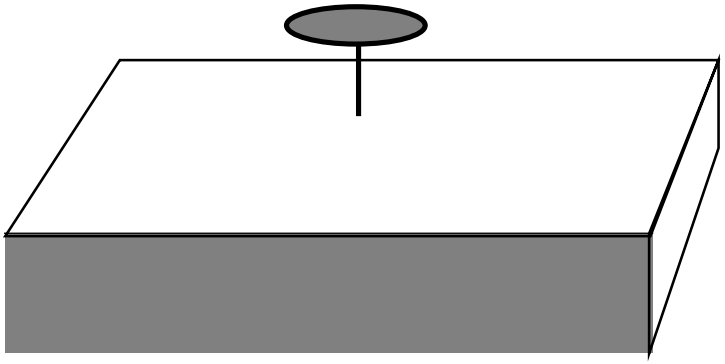
In the special case of an L-antenna, however, we can derive an approximate formula by making the following assumption (windbreaker required here): Pretend that the current distribution along a straight conductor is only moderately perturbed when the antenna is bent into an L-shape. If this cheesy assumption holds, then we have already derived the relevant formula:

$$R_r \approx 40\pi^2 \left(\frac{l+d}{\lambda} \right)^2, \tag{13}$$

where l and d are as defined in the figure, and the total length is assumed short compared to a wavelength. This equation is so approximate that one should expect the need to trim the antenna to the proper length. However, it is a reasonable guide to establish rough dimensions for an initial design.

If the primary value of the horizontal segment is in boosting capacitance, then further improvements might be enabled by using more segments. Commonly 2 (for a T-antenna), 3 and 4 horizontal conductors are used, symmetrically arranged about the vertical portion. The capacity hat may be considered the limit of using an infinite number of radial conductors:

FIGURE 6. Antenna with capacity hat



Other capacitive structures, such as spheres and spheroids, have also been used in place of the flat disk shown in Figure 6.

4.4 Inductively loaded dipole

After all the discussion about how radiation is generally insignificant until conductor dimensions are some reasonable fraction of a wavelength, it may be somewhat surprising that the signal power available from a dipole antenna at any single frequency is actually independent of its length. This invariance can be understood by observing first that shorter dipoles deliver lower *voltages*. To first order, one may take the open-circuit voltage as equal to the product of antenna length and received electric field strength, so voltage scales linearly with length for short dipoles (up to a point). At the same time, we've seen that the radiation resistance varies as the square of length. Hence the ratio of voltage squared to resistance is independent of length. As the dipole length diminishes, the lower voltage is delivered from lower Thévenin resistances (the radiation resistances), such that the available power remains constant. For a lossless monopole, for example, the available power is

$$P_{av} = \frac{(E_{pk} l)^2 / 8}{40\pi^2 \left(\frac{l}{\lambda}\right)^2} = \frac{(E_{pk}\lambda)^2}{320\pi^2}. \tag{14}$$

Clearly, the available power is independent of length, and instead depends only on the field strength and wavelength. Thus, for the lossless dipole assumed, theory says that we could make antennas out of infinitesimally short segments. This conclusion is seemingly at odds with ordinary experience, where radiation from ordinary wires is routinely ignored with impunity in low frequency circuit analysis, and where AM radio stations use antennas of such a size that they must be supported by very tall towers. The resolution to this apparent paradox is that the radiation resistance forms a voltage divider with the Thévenin equivalent resistance of the driving source, augmented by ever present resistive losses in any circuit. In fact it is precisely this implicit impedance mismatch that allows us to glibly ignore radiation from short wires used as interconnect. Any wire is capable of radiating at any time, but if it's short, the impedance mismatch is typically so great that very little energy is delivered to the radiation resistance. *That's* how a wire knows when and when not to act as an antenna in ordinary circuits.

Suppose, though, that we were able to avoid this impedance mismatch. After all, impedance transformers are readily designed. Could we then make antennas arbitrarily short? The answer is a qualified Yes. One qualification can be appreciated after recognizing that bandwidth and normalized antenna length are actually coupled. Because short dipoles have a capacitive reactive component, addition of a suitable inductance will permit the antenna circuit to be brought into resonance at a given desired frequency of operation. Electrically speaking, the antenna acts longer insofar as the disappearance of a reactive term is concerned. These loading inductances are usually placed either at the base of the dipole (i.e., at the feedpoint), or near the center of the dipole. However, as the antenna shrinks, so does its capacitance. To maintain resonance, the compensating (loading) inductance must increase. Recalling that the *Q* of a series resonant circuit is

$$Q = \frac{\sqrt{L/C}}{R}, \tag{15}$$

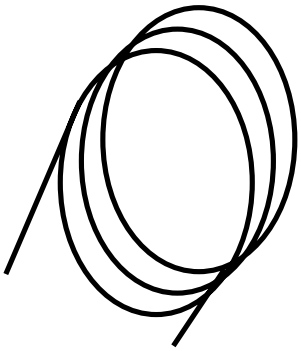
it should be clear that Q increases as the antenna shortens (assuming no losses), because both inductance and capacitance are roughly proportional to length, and the radiation resistance is proportional to the square of the length. The bandwidth is therefore proportional to the first power of length. As a result allowable reductions in antenna length are limited by the desired communication bandwidth. Furthermore, the narrower the bandwidth, the more sensitive the antenna's center frequency to the proximity of objects. Purposeful addition of series resistance to mitigate this sensitivity and also improve bandwidth is accompanied by an unfortunate increase in loss. Even if no additional resistance is provided intentionally, there is always some loss. If efficiency is to remain high, the additional series resistance representing this loss must be small compared with the radiation resistance. To underscore the practical difficulties involved, consider a monopole antenna that is 1% of a wavelength. The radiation resistance is then about 0.04Ω. Needless to say, it is exceedingly difficult to arrange for all RF losses to be small compared to resistances of forty milliohms! The fundamental tradeoff between efficiency and bandwidth thus tightly constrains the practical extent to which a dipole may be shortened. The coupling among bandwidth, normalized length and efficiency drives most antenna designs to at least as long as about 10% of wavelength. Practical dipole antennas are rarely much shorter than this value, except for applications where the available signal power is so large that inefficient antennas are acceptable.

Occasionally, one will encounter antennas that employ both capacitive and inductive loading (e.g., capacity hat plus loading coil). The additional degree of freedom can permit one to relax the tradeoff to a certain extent.

4.5 Magnetic loop

The dual of an electric dipole is the magnetic dipole formed by a loop of current. Just as the dipole antenna is sensitive primarily to the electric field, the loop antenna is sensitive mainly to the magnetic field component of an incoming wave. We'll see momentarily that this duality makes the loop antenna attractive in many situations where the dipole antenna suffers from serious problems. In particular, at low frequencies, a loop antenna design is frequently more practical than its electric dipole counterpart, explaining why loop antennas are almost universally used in portable AM radios and in many pagers, for example.

FIGURE 7. Loop antenna



The following equation for the effective radiation resistance of a circular loop antenna assumes that the diameter is very short compared to a wavelength, and that no magnetic materials are used.[18]

$$R_{rad} \approx \frac{80\pi^5 n^2 d^2}{l^4}, \tag{16}$$

where n, d, and l are the number of turns, loop diameter and loop length, respectively.

Just as the short electric dipole antenna produces a net capacitive reactance, the magnetic loop antenna has a net inductive reactance. Wheeler's famous formula can be used to predict the inductance of an air-core loop:

$$L \approx \frac{10\pi\mu_o n^2 r^2}{9r + 10l}, \tag{17}$$

which assumes dimensions and inductance in SI units, unlike Wheeler's original formulation, which uses dimensions in inches.

It is important to take note of a new degree of freedom not present in the dipole case: one can add more turns to increase radiation resistance. This improvement comes about in the same way as does the impedance transformation of a conventional transformer. The changing magnetic field of the incoming wave induces the same voltage in each turn, so we get n times the per-turn voltage at the antenna terminals. Since energy must be conserved, the current must drop by this same factor n, so that the resistance (the ratio of voltage to current, says Professor Ohm) increases by n^2.

We may now appreciate how the loop antenna can solve the thorny problem of AM radio reception. Signals at the lower end of the AM band possess a wavelength of almost 600 meters. The maximum allowable dimensions of any portable device will necessarily be an

18. H. A. Wheeler, *Fundamental Limitations of Small Antennas*, Proc. IRE, Dec. 1947, pp. 1479-1488. If magnetic materials are used, then the radiation resistance is multiplied by a factor that is a function of the permeability and the geometry.

absurdly small fraction of this wavelength. A standard dipole antenna of any human-sized dimensions would thus have an infinitesimal radiation resistance, making efficient operation practically impossible. The loop antenna offers a welcome alternative. It is chiefly for this reason that loop antennas are the only type of antenna used in portable AM radios. Further improvements are provided by winding the antenna around a ferrite core, whose large permeability concentrates the magnetic field. These “loopstick” antennas dominate portable applications up to frequencies where the lossiness of ferrites negates their usefulness (perhaps as high as the VHF range). Loop antennas are also the choice in pagers, where the desire for a very small form factor makes it difficult to realize an efficient dipole. The loop is conveniently shaped as a rectangle and mounted inside the case of the pager.

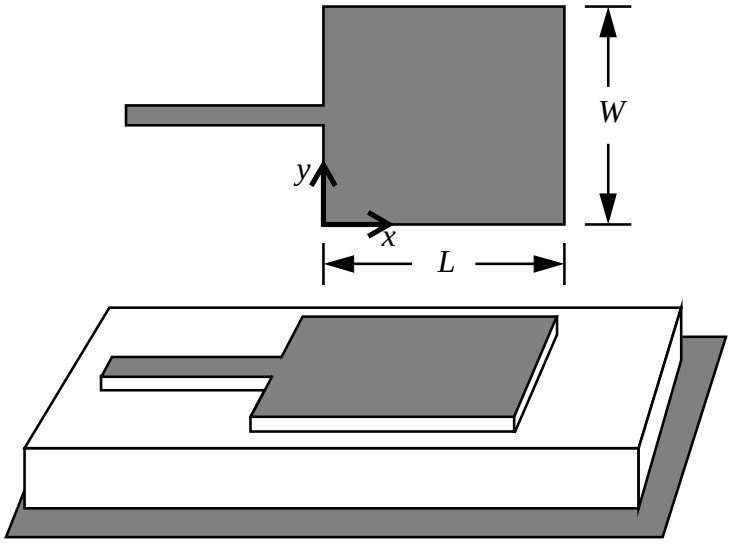
5.0 The Microstrip Patch Antenna

We’ve seen that whenever a conductor is an appreciable fraction of a wavelength, radiation becomes practical. This effect is not always wanted; for example, radiation losses increase the attenuation of microstrip lines. While undesirable in that context, such radiation is of course precisely what is required to make antennas. When built out of microstrip, these radiators are known as patch antennas. They have become extremely popular because of their planar nature, making them amenable to inexpensive batch fabrication, just as any other printed circuit. Despite a number of important limitations (*excessive Q* or, equivalently, excessively narrow bandwidth, and a tendency to radiate from the feed), the convenience and compactness more than compensate in many applications.

To first order, the patch antenna can be considered the limiting case of connecting a planar array of thin dipoles in parallel so that they form a sheet. As such, the primary radiation is normal to the surface of the patch. The precise nature of the radiation pattern can be adjusted within fairly wide limits by controlling how one feeds the antenna. Typically, patches are fed at one end, at the center of an edge (see figure). However, one may also use off-center feeds (offset feeds) to excite other than linear polarizations. This ability is highly valuable, for many microwave communications systems employ polarizations to provide a measure of multipath mitigation.[19]

19. Reflection off of an object reverses the sense of polarization, changing a counterclockwise polarization into a clockwise one, for example. Using an antenna that selectively rejects one of these thus reduces a communications link’s susceptibility to troublesome reflections.

FIGURE 8. Halfwave patch antenna (conductor pattern and perspective view)



For the antenna in Figure 8, assume that the length L is chosen equal to a half wavelength. In that case, the current is zero at $x = 0$ and $x = L$, with a maximum at $L/2$, as in a classic dipole antenna. At the same time, the voltage is a minimum at $L/2$, and a maximum at the source and far end, again just as in a classic dipole.

One important characteristic of a patch antenna is its relatively narrow bandwidth (typically ranging from a fraction of a percent to a few percent). This quality is a double-edged sword; it endows the patch with an ability to filter off-frequency signals, but it also demands more of manufacturing precision and material stability. The variability of FR4 is large enough that a certain amount of cutting-and-trying is to be expected, generally limiting its use to prototyping and hobbyist applications.

Numerous design equations have appeared in the literature, spanning a broad range of complexity. In the interest of preserving the maximum level of intuitive value consistent with usefulness, the equations presented here are simple. They are typically in error by an amount similar to that caused by parameter variation in general-purpose PC board materials and manufacture. That's just a fancy way of saying that one should still expect the need to perform some trimming, whether you use the following equations or not.

As suggested earlier, the classic patch antenna is designed as a half-wave radiator, so its electrical length is chosen equal to a half wavelength:

$$L_{eff} = \frac{\lambda}{2}. \tag{18}$$

In relating electrical and physical lengths, it's important to consider both fringing field and the effective dielectric constant:

$$L_{eff} \approx \sqrt{\varepsilon_{r,\,eff}}\left(L + 2\frac{H}{2}\right) = \sqrt{\varepsilon_{r,\,eff}}\,(L + H)\,, \tag{19}$$

where H is the thickness of the dielectric, and the length correction per edge, $H/2$, is the same as derived in the appendix of the chapter on microstrip.

The effective dielectric constant is given by

$$\varepsilon_{r,\,eff} \approx 1 + 0.63 \cdot (\varepsilon_r - 1) \cdot \left(\frac{W}{H}\right)^{0.1255}, (W/H > 0.6) \quad , \tag{20}$$

which is the same formula as used for ordinary microstrip lines (as is the correction for fringing in Eqn. 19). Marginally easier to remember is the following alternative approximation:

$$\varepsilon_{r,\,eff} \approx 1 + \frac{5}{8} \cdot (\varepsilon_r - 1) \cdot \left(\frac{W}{H}\right)^{\frac{1}{8}}, (W/H > 0.6) \quad . \tag{21}$$

Since the width W of a typical patch antenna is so much greater than the dielectric thickness H, the effective dielectric constant is usually quite close to the dielectric constant of the material (say, only 5% below it). For that reason, design formulas presented in many references do not make a distinction between these two dielectric constants.

We need to perform a similar accommodation of fringing effects on the effective width:

$$W_{eff} \approx \sqrt{\varepsilon_{r,\,eff}}\left(W + 2\frac{H}{2}\right) = \sqrt{\varepsilon_{r,\,eff}}\,(W + H) \, . \tag{22}$$

Continuing with our design equations, we obtain:

$$f_0 \approx \frac{1.5\times 10^8}{L_{eff}}, \tag{23}$$

and

$$Z_0 \approx \frac{90\left(\varepsilon_{r,\,eff}\dfrac{(L+H)}{(W+H)}\right)^2}{\varepsilon_{r,\,eff} - 1} . \tag{24}$$

Notice that one way to control the impedance of a patch antenna is to choose an appropriate ratio of length to width. For patches made on FR4 with an effective relative dielectric constant of 4.2, this formula says that a ~50Ω feedpoint impedance results from a W/L ratio of about 3, whereas a square patch typically presents an impedance of about 500Ω.

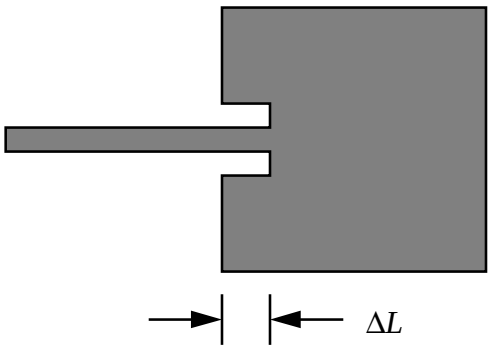
Because the length is set by the desired center frequency, increasing the width to provide a lower driving point impedance sometimes results in patches of inconveniently large size. In those situations, one may use alternative impedance transformation techniques. A classical one is to interpose a quarter-wavelength segment of line between the source and the

antenna. If the transforming line's characteristic impedance is made equal to the geometric mean of the source and load impedances, a match results.

As a specific numerical example, suppose we need to design a 50Ω patch antenna for use in a portable application in the 2.5GHz ISM (industrial, scientific and medical) frequency band. Using FR4, a rectangular patch would have dimensions of about 27mm by 80mm. The length of just a bit over an inch is reasonable, but the width isn't quite compatible with most portable form factors. Suppose that we choose a square patch instead, whose impedance is 500Ω. A 160Ω quarter-wavelength line would perform the necessary transformation. Realizing such an impedance generally requires a very narrow line, and manufacturing tolerances are consequently critical in such a case.

Yet another impedance transformer option is available with patches (indeed, with any resonant antenna). Because of the standing wave set up in the antenna, voltages and currents vary along the patch. In the half-wave case we've been studying, the boundary conditions force the current to be a minimum at the feedpoint and at the far end of the patch, with a maximum in the middle. At the same time, the voltage is a minimum in the middle, and at a maximum at the source and far end. The impedance, being the ratio of voltage to current, therefore varies along the antenna, from a minimum at the normal feedpoint, to a maximum at the middle of the patch. One may exploit this impedance variation by using an inset feed, as shown in the following figure:

FIGURE 9. Halfwave patch antenna with impedance-transforming inset feed (top view)



To a first approximation, the standing waves (voltage and current) vary sinusoidally along the length of the patch (these assumptions are the same as those used in deriving the radiation resistance of an ordinary dipole). The current is nearly zero at both ends of the patch, increasing roughly sinusoidally as one moves toward the center. At the same time, the voltage is a peak at the ends, sinusoidally decaying toward zero in the center. Therefore, as one moves the feedpoint toward the center, the ratio of voltage to current varies approximately quadratically because voltage decreases sinusoidally at the same rate that the current increases. The impedance therefore gets multiplied by the following factor:

$$Z = (Z_{edge}) \left(\cos \pi \frac{\Delta L}{L} \right)^2, \tag{25}$$

where Z_{edge} is the driving point impedance in the absence of an inset feed.

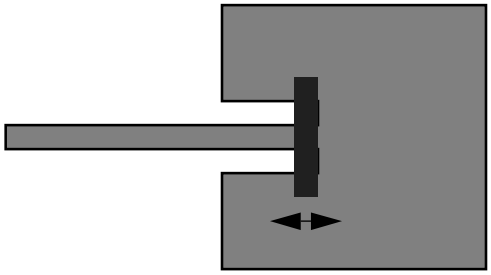
For our example, we need to transform downward by the comparatively large factor of about 10.8, implying that

$$\Delta L = \frac{L}{\pi}\text{acos}\left(\frac{1}{\sqrt{10.8}}\right) \approx 0.4L \, . \tag{26}$$

Note that the inset is nearly all the way to the middle. Because the change in voltage with distance is large near the middle[20], the precise value of the impedance is a sensitive function of distance in the vicinity of this inset's location. That, plus the uncertainty inherent in our approximations, means again that some empirical adjustment will probably be necessary to obtain the correct impedance.

One convenient method for trimming such an inset patch is first to use a deeper-than-nominal inset. Then, upward impedance adjustments are easily effected by placing a shorting strip across some portion of the inset:

FIGURE 10. Adjustment method for inset patch (top view)



This particular method avoids the need for precision cutting, and also facilitates multiple iterations. Soldering (and unsoldering) a piece of copper foil tape is much easier than gouging out segments of copper cladding.

A disadvantage of the inset feed is that it perturbs the field distributions, with the amount increasing with the depth of the inset. The three impedance matching methods (controlling *W*/*L*, quarter-wave transformer and inset feed) provide important degrees of freedom for trading off parameters of interest. For example, one may use a rectangular patch with dimensions that produce a 200Ω impedance, combine it with an inset feed to drop it to 100Ω, then finish with a quarterwave match (whose line impedance is 71Ω) to get to 50Ω. Since the impedance transformations at each step along the way involve relatively small ratios, a more practical, robust design results.[21]

20. The voltage is a minimum in the center, but its spatial derivative is a *maximum* there.

21. Of course, one could also use a sequence of quarterwave transformers, or some other variations. There are many ways to accomplish the needed transformation, and the reader is invited to explore alternatives independently.

6.0 Summary

We've seen that radiation is fundamentally the result of the finite propagation speed of light. The need for a reasonably large phase shift to produce a reasonably high radiation resistance explains why real antennas are a reasonable fraction of a wavelength in extent, at minimum.

Not only are elementary dipole antennas (both balanced and grounded) quite commonly used, they serve as an important basis for understanding more complex antennas. Short dipoles have low radiation resistances and are primarily capacitive. Capacity hats can be used to increase radiation resistance, and inductances can be used to tune out any capacitance. Such measures are effective up to a limit imposed by the need to provide a given minimum bandwidth or efficiency, and to produce an antenna whose characteristics are not overly sensitive to small changes in dimensions or environmental conditions. The tradeoffs are such that antennas much shorter than about a tenth of a wavelength are frequently regarded as unsatisfactory.

The magnetic loop antenna may be viewed as the dual of the electric dipole. Unlike the dipole, the radiation resistance depends on the number of turns, endowing it with an additional degree of freedom that makes it possible to realize compact antennas.

Finally the patch antenna can be considered a continuous parallel connection of an infinite number of infinitesimally thin dipoles. Although its excessive Q is a definite disadvantage in many situations its other attributes, such as amenability to batch manufacturing, often more than compensate.

Filters

1.0 Introduction

The subject of filter design is so vast that we have to abandon any hope of doing justice to it as a subset of a textbook. Indeed, even though we have chosen to present this material over two chapters, the limited aim here is to focus on important qualitative ideas and practical information about filters, instead of attempt a comprehensive review of all possible filter types and supply complete mathematical details of their underlying theory. For those interested in the rigor that we will tragically neglect, we will be sure to provide pointers to the relevant literature. And for those who would rather ignore the small amount of rigor that we do provide, the reader is invited to skip directly to the appendices, which summarize filter design information in "cookbook" form.

Although our planar focus would normally imply a discussion limited to microstrip implementations, many such filters derive directly from lower frequency lumped prototypes. Because so many key concepts may be understood by studying those prototypes, we will follow a roughly historical path and begin with a discussion of lumped filter design. It is definitely the case that certain fundamental insights are universal, and it is these that will be emphasized in this chapter, despite differences in implementation details between lumped and distributed realizations.

Only passive filters will be considered here, partly to limit the length of the chapter to something manageable. Another reason is that, compared to passive filters, active filters generally suffer from higher noise and nonlinearity, limited operational frequency range, higher power consumption, and relatively high sensitivity to parameter variations, particularly at the GHz frequencies with which we are concerned in this textbook.

2.0 Background

2.1 A quick history

The use of frequency selective circuits certainly dates back at least to the earliest research on electromagnetic waves. In his classic experiments of 1887-1888 Hertz himself used dipole and loop antennas (ring resonators) to clean up the spectrum generated by his spark gap apparatus and thereby impart a small measure of selectivity to his primitive receivers. Wireless pioneer Sir Oliver Lodge of the U.K. coined the term "syntony" to describe the action of tuned circuits, showing a conscious appreciation of the value of such tuning, despite the hopelessly broadband nature of spark signals.[1] At nearly the same time, Nikola Tesla and Guglielmo Marconi developed tuned circuits (Marconi's patent #7777 was so

1. See H. Aitken's excellent book, *Syntony and Spark*, Princeton, 1987, for a technically detailed and fascinating account of early work in wireless.

valuable that it became the subject of bitter and protracted litigation)[2] for the specific purpose of rejecting unwanted signals, anticipating the advent of sinusoidal carrier based communications.

Despite that foundation, however, modern filter theory does not trace directly back to those early efforts in wireless. Rather the roots go back even further in time: it is research into the properties of transmission lines for telegraphy and telephony that primarily inform early filter theory. In 1854 William Thomson (who would later become Lord Kelvin), carried out the first analysis of a transmission line, considering only the line's distributed resistance and capacitance. His work, inspired by what was to be the 3000-kilometer Atlantic Cable Project, established a relationship between practical transmission rates and line parameters. A bit over 20 years later, Oliver Heaviside and others augmented Kelvin's analysis by including distributed inductance, thereby extending greatly the frequency range over which transmission line behavior could be described accurately.[3] Following up on one particular implication of Heaviside's work, both George Ashley Campbell of the American Bell Company and Michael Idvorsky Pupin of Columbia University suggested around 1900 the insertion of lumped inductances at regularly spaced intervals along telephone transmission lines to reduce dispersion (the smearing out of pulses).[4] This suggestion is relevant to the filter story because Heaviside recognized that a lumped line differs from a continuous one in possessing a definite cutoff frequency. Campbell and Pupin provided design guidelines for guaranteeing a certain minimum bandwidth.[5]

In true engineering fashion, the apparent liability of a lumped line's limited bandwidth was quickly turned into an asset, and thus was established the main evolutionary branch of filter design. The first published formalism is Campbell's, whose classic 1922 paper describes in fuller detail ideas he had developed and patented during WWI.[6] Karl Willy Wagner also developed these ideas at about the same time, but German military authorities delayed publication, giving Campbell priority.[7] It is now acknowledged that credit should be shared by these two pioneers, who independently and nearly simultaneously hit upon the same great idea.

---

2. The U.S. Supreme Court eventually ruled it invalid (in 1943) because of prior work by Lodge, Tesla and others.

3. For additional background on this story, see Paul J. Nahin's excellent book, *Oliver Heaviside: Sage in Solitude*, IEEE Press, 1987.

4. As with many key ideas of great commercial import, a legal battle erupted over this one. It is a matter of record that the Bell System was already experimenting with loading coils developed by Campbell well before publication of Pupin's 1900 paper. Pupin's self-promotional abilities were superior, though, and he was able to obtain a patent nonetheless. He eventually earned royalties of over $400,000 from Campbell's employer (at a time when there was no U.S. income tax) for his "invention." To add to the insult, Pupin's Pulitzer-prize winning autobiography of 1924 shamefully fails to acknowledge Campbell and Heaviside.

5. A. T. Starr, *Electric Circuits and Wave Filters*, 2nd ed., Pitman and Sons, 1948.

6. G. A. Campbell, "Physical Theory of the Electric Wave-Filter," *Bell System Technical Journal*, vol. 1, no. 2, pp. 1-32, Nov. 1922. See also his U. S. Patent #1,227,113, May 22, 1917.

7. "Spulen- und Kondensatorleitungen" (Inductor and Capacitor Lines), *Archiv für Electrotechnik*, vol. 8, July 1919.

Campbell's colleague, Otto J. Zobel, published a much-referenced extension of Campbell's work, but which was still limited to filters derived from transmission line ideas.[8] In the developments of subsequent decades one sees an evolving understanding of how closely one may approach in practice the theoretical ideal of a perfectly flat passband, constant group delay, and an infinitely steep transition to an infinitely attenuating stopband. Conscious acknowledgment that this theoretical ideal is unattainable leads to the important idea that one must settle for approximations. Some of the more important, practical and well-defined of these approximations are the Butterworth, Chebyshev and Cauer (elliptical) filter types we'll study in this chapter.

Shortly after WWII, the subject of filter design advanced at an accelerated pace. Investigation into methods for accommodating finite-*Q* elements in lumped filters offered hope for improved predictability and accuracy. In the microwave domain, filter topologies based directly on lumped prototypes came to be supplemented by ones that exploit, rather than ignore, distributed effects. Many of these are readily implemented in microstrip form, and are the ultimate focus of this chapter.

The advent of transistors assured that the size of active devices no longer dominated that of a circuit. Numerous active filter topologies evolved to respond to a growing demand for miniaturization, replacing bulky passive inductor-capacitor circuits in many instances. Aside from enabling dramatic size reductions, some active filters are also electronically tunable. However, these attributes do come at a price: active filters consume power, suffer from nonlinearity and noise, and possess diminished upper operational frequencies because of the need to realize gain elements with well-controlled characteristics at high frequencies. These tradeoffs become increasingly serious as microwave frequencies are approached. This statement should not be taken to mean that microwave active filters can never be made to work well enough for some applications (because successful examples certainly abound), but it remains true that the best filters at such frequencies continue to be passive implementations. It is for this reason that this chapter considers passive filters exclusively.

The arrival of transistors also coincided with (and helped drive) a rapidly decreasing cost of computation.[9] No longer limited to considering only straightforward analytical solutions, theorists were free to pose the filter approximation problem much more generally, e.g., "Place the poles and zeros of a network to minimize the mean-square error (or maximum error, or some other performance metric) in a particular frequency interval, relative to an ideal response template." Of great practical importance is that a numerical approach readily accommodates lossy inductors and capacitors, something that is difficult with earlier analytical approaches. The resulting filters are optimum in the sense that one cannot do better (as evaluated by whatever design criteria were imposed in the first place) for a given filter order. The tradeoff is that the resulting design may not be as easily understood

8. O. J. Zobel, "Theory and Design of Uniform and Composite Electric Wave-filters," *Bell System Technical Journal*, vol. 2, no. 1, pp. 1-46, Jan. 1923.

9. Regrettably, space limitations force us to neglect here the fascinating story of electrolytic tanks and other analog computers used to design filters based on potential theory.

as those based on analytical approaches. In many cases, element values are best obtained from tables that summarize the results of extensive computations.

The same philosophical approach of numerical optimization is also how most modern microwave filters are designed. And again, solutions for the more complex types are best extracted from tables. The main purpose here is to provide an intuitive explanation for how these filters work, leaving many of the mathematical details to published theoretical treatments.

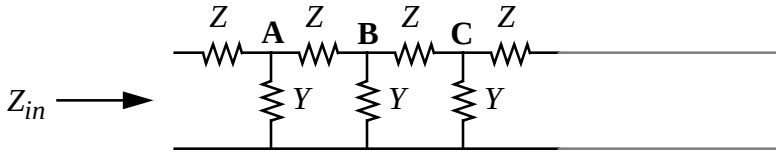
3.0 Filters from Transmission Lines

We start with the "electric wave filters" of Campbell, Zobel and Wagner. As mentioned, these derive from lumped approximations to transmission lines, so we begin by examining such "artificial" lines to see how a limited bandwidth arises.

3.1 Constant-*k* filters

For convenience, we repeat here some of the calculations from the chapter on distributed systems. Recall that we first consider the driving point impedance, Z_{in}, of the following infinite ladder network:

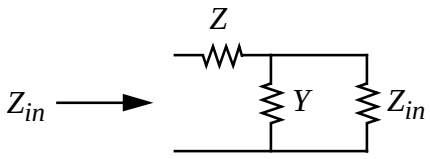
FIGURE 1. Infinite ladder network as artificial line



In this drawing, the resistor symbols represent generalized impedances (Z) and admittances (Y).

To simplify the derivation, it is helpful to note that the impedance looking to the right of point **A** also equals Z_{in}, this being an infinite network. We can then collapse the infinite network into a much simpler finite one:

FIGURE 2. Conversion of infinite line into finite network



Solving for the Z_{in} of this simple network yields:

$$Z_{in} = \frac{Z \pm \sqrt{Z^2 + 4\,(Z/Y)}}{2} = \frac{Z}{2}\left[1 \pm \sqrt{1 + \frac{4}{ZY}}\right], \tag{1}$$

where one would generally disallow negative values and thus choose only the sum solution.

As a specific (but typical) case, consider a low-pass filter in which $Y = j\omega C$ and $Z = j\omega L$. Then, the input impedance of the infinite artificial line is:

$$Z_{in} = \frac{j\omega L}{2}\left[1 \pm \sqrt{1 - \frac{4}{\omega^2 LC}}\right]. \tag{2}$$

At very low frequencies, the factor under the radical is negative and large in magnitude, making the term within the brackets almost purely imaginary. The overall Z_{in} in that frequency range is therefore largely real, with

$$Z_{in} \approx \sqrt{\frac{Z}{Y}} = \sqrt{\frac{L}{C}} = k\,. \tag{3}$$

Because the ratio Z/Y is a constant here, such filters are often known as constant-k filters.[10]

As long as the input impedance has a real component, nonzero average power will couple into the line from the source. Above some particular frequency, however, the input impedance becomes purely imaginary, as can be seen from inspection of Eqn. 2. Under this condition, no real power can be delivered to the network, and the filter thus attenuates heavily.[11] For self-evident reasons the frequency at which the input impedance becomes purely imaginary is called the cutoff frequency which, for this low-pass filter example, is given by:

$$\omega_h = \frac{2}{\sqrt{LC}}. \tag{4}$$

Any practical filter must employ a finite number of sections, of course, leading to the question of the relevancy of any analysis that assumes an infinite number of sections. Intuitively, it seems reasonable that a "sufficiently large" number of sections would lead to acceptable agreement. Based on lumped network theory, we also expect the ultimate rate of rolloff to be determined by the filter order, and hence by the number of sections to which the network is truncated (we'll have more to say on this subject later). The greater

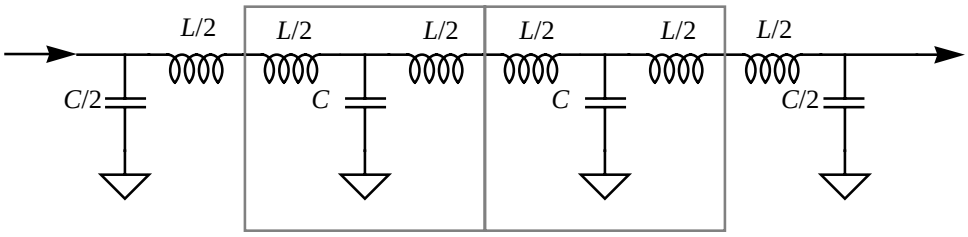
10. Campbell used the symbol *k* in precisely this context, but it was Zobel (op. cit.) who apparently first used the actual term "constant-*k*."

11. Attenuation without dissipative elements might initially seem intuitively unpalatable. However, consider that a filter might also operate by *reflecting* energy, rather than by dissipating it. That is, a filter can function by producing a purposeful impedance mismatch over some band of frequencies. In fact, many filter design approaches are based directly on manipulation of the reflection coefficient as a function of frequency.

the number of sections, the greater the rate of rolloff. As we'll see, there is also some (but practically limited) flexibility in the choice of *Z* and *Y*, permitting a certain level of trade-off among passband, transition band, and stopband characteristics. However, it remains true that one limitation of filters based on artificial-line concepts is the inability to specify these characteristics in detail, if at all. Note, for example, the conspicuous absence of any discussion about how the filter behaves near cutoff. We don't know if the transition from passband to stopband is gradual or abrupt, monotonic or oscillatory. We also don't know the precise shape of the passband. Finally, we don't have any guide how to modify the transition shape should we find it unsatisfactory. As we'll see, these shortcomings lead us to consider other filter design approaches.

Once the filter order is chosen (by whatever means), the next problem is one of termination. Note that the foregoing analysis assumes that the filter is terminated in an impedance that behaves as described by Eqn. 2. That is, our putative finite filter must be terminated in the impedance produced by the prototype *infinite* ladder network: It must have a real impedance at low frequencies, then become purely imaginary above the cutoff frequency. Stated another way, rigorous satisfaction of the criteria implied by Eqn. 2 absurdly requires that we supply a load element which itself is the filter we desire! We should therefore not be too surprised to discover that a practical realization involves compromises, all intimately related to the hopeless task of mimicking the impedance behavior of an infinite structure with a finite one. For example the near-universal choice is to terminate the following with a simple resistance *R* equal to *k*:

FIGURE 3. Low-pass constant-*k* filter example using two cascaded T-sections



A source with a Thévenin resistance also of value *k* is assumed to drive this filter. Note that this example uses two complete T-sections (shown in the boundaries), with a *half*-section placed on each end. Termination in half sections is the traditional way to construct such filters. The series-connected inductors, shown individually to identify clearly the separate contributions of the unit T-sections, are combined into a single inductance in practice. Alternatively one may implement the filter with π-sections. With those, one uses terminating half-sections that are mirror images of the ones shown. The choice of which implementation to employ is often determined in practice by the nature of the parasitics that dominate the input and output interfaces. If these parasitics are primarily capacitive in nature, the T-section implementation shown is favored, since the parasitics may be absorbed into the capacitances at the ends of the filter. Similarly, inductive parasitics are most readily accommodated by a filter using internal π-sections.

The design equations for this filter are readily derived from combining Eqn. 3 and Eqn. 4:

$$C = \left(\frac{2}{\omega_h}\right)\frac{1}{R}, \tag{5}$$

and

$$L = \left(\frac{2}{\omega_h}\right)R. \tag{6}$$

Thus, once one specifies the characteristic impedance, R, the desired cutoff frequency, and the total number of sections, the filter design is complete.

Regrettably, deducing the number of sections required is a bit of a cut and try affair in practice. There are equations that can provide guidance, but they are either cumbersome or inaccurate enough that one typically increases the number of sections until simulations reveal that the filter behaves as desired. Furthermore the unsophisticated termination of a simple resistance leads to degradation of important filter characteristics, often resulting in a hard-to-predict insertion loss and passband flatness, as well as reduced stopband attenuation (relative to predictions based on true, infinite-length lines). These difficulties are apparent from an inspection of the following table, which shows the attenuation at the cutoff frequency, as well as the –3dB and –6dB bandwidths (expressed as a fraction of the cutoff frequency), of constant-*k* filters (both T- and π-implementations) as a function of order. In the table, *n* is the number of complete T- (or π-) sections in the central core of the filter. The filter order is therefore $2n + 3$.

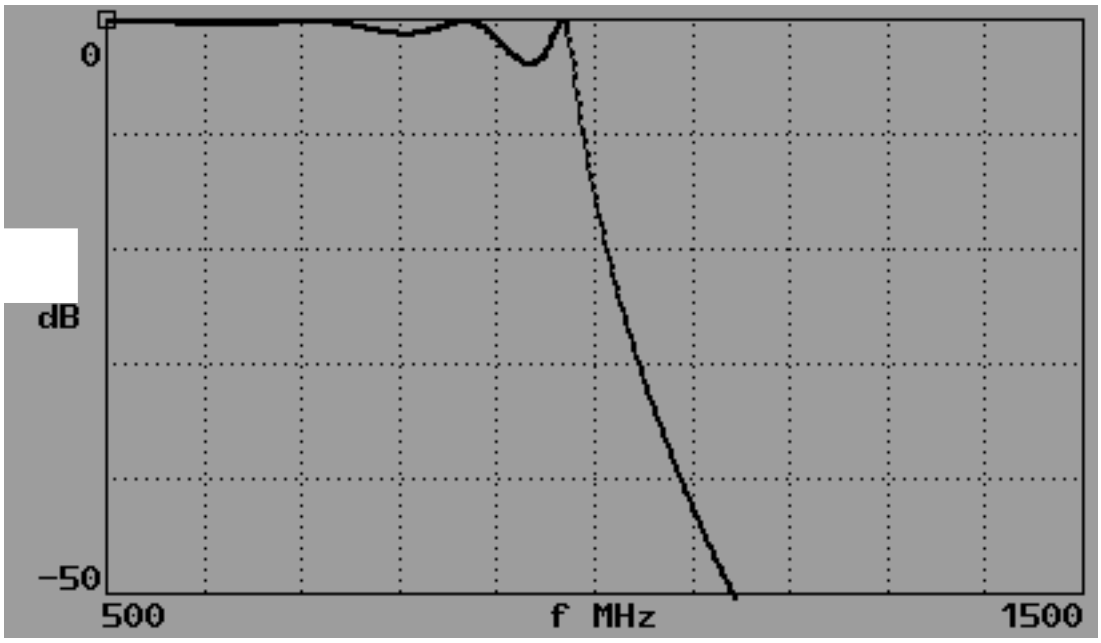
TABLE 1. Characteristics of ideal constant-*k* filters

n	***Attenuation at cutoff frequency (dB)***	***Normalized –3dB Bandwidth***	***Normalized –6dB Bandwidth***	***Normalized –60dB Bandwidth***	***Normalized –10dB S_{11} Bandwidth***
0	3.0	1.000	1.201	10.000	0.693
1	7.0	0.911	0.980	3.050	0.810
2	10.0	0.934	0.963	1.887	0.695
3	12.3	0.954	0.969	1.486	0.773
4	14.2	0.967	0.976	1.302	0.696
5	15.7	0.976	0.981	1.203	0.756

Note that the attenuation at the nominal cutoff frequency, as well as the bandwidth, are both dependent on the number of filter sections. Further note that the cutoff frequency (as computed by Eqn. 4) equals the –3dB bandwidth only for $n = 0$, and is as much as 10% beyond the –3dB bandwidth in the worst case. In critical applications, the cutoff frequency target may have to be altered accordingly to achieve a specified bandwidth.

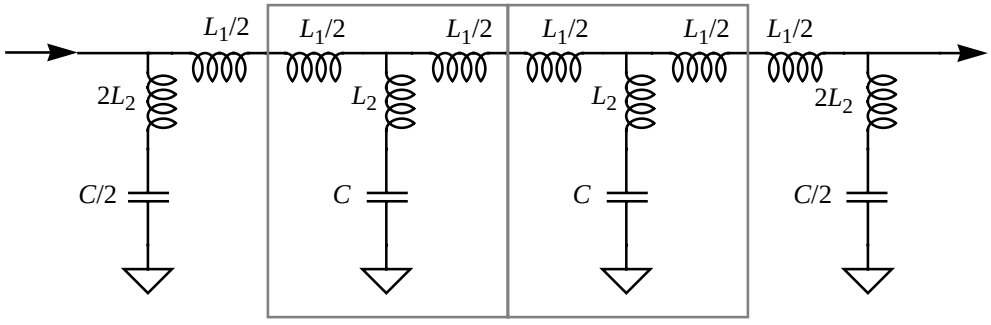
The following figure is a frequency response plot for a constant-*k* filter which consists of five full sections, and a terminating half-section on each end. Note that the frequency axis is linear, not logarithmic:

FIGURE 4. Response of six-section low pass constant-*k* filter (*n* = 5)



Aside from the ripple evident in the figure, it is also unfortunate that the bandwidth over which the return loss exceeds 10dB is typically only ~70-80% of the cutoff frequency. A considerable improvement in performance is possible by using filter sections whose impedance behavior better approximates a constant resistance over a broader frequency range. One example, developed by Zobel, uses "*m*-derived" networks, either as terminating structures, or as filter sections:

FIGURE 5. Low-pass *m*-derived filter using cascaded T-sections



As with the prototype constant-*k* filter of Figure 5, this structure is both driven and terminated with a resistance of value *k* ohms. The *m*-derived filter, which itself is a constant-*k* structure, is best understood by noting that the prototype constant-*k* filter previously analyzed has a response that generally attenuates more strongly as the cutoff frequency ω_1 is

approached. At small fractions of the cutoff frequency, the response is fairly flat, so it should seem reasonable that increasing the cutoff frequency to some value ω_2 should produce a more constant response within the original bandwidth ω_1. The first step in designing an *m*-derived filter, then, consists simply of increasing the cutoff frequency of a prototype constant-*k* filter. In the absence of inductor L_2, we see that scaling the values of L_1 and C each, say, by a factor m (with m ranging from 0 to 1) increases the cutoff frequency by a factor of $1/m$, from a value ω_1, to $\omega_2 = \omega_1/m$. The characteristic impedance remains unchanged at k because the *ratio* of L_1 to C is unaffected by this scaling.

Now to restore the original cutoff frequency, add an inductance L_2 to produce a series resonance with C. At the resonant frequency, this series arm presents a short circuit, creating a notch in the filter's transmission. If this notch is placed at the right frequency (just a bit above the desired cutoff frequency), the filter's cutoff frequency can be brought back down to ω_1. However, be aware that the filter response does pop back up above the notch frequency (where the resonant branch then looks like a simple inductance). This characteristic needs to be taken into account when using the *m*-derived filter.

An alternative to a series resonance in the shunt arm of each filter section is a parallel resonance in the series arm(s) of a filter section. Both types of *m*-derived filters will provide the same behavior. The choice of topology in practice is often determined by which implementation uses more easily realized components, or which more gracefully accommodates parasitic elements.

Following a procedure exactly analogous to that used in determining the cutoff frequency of ordinary constant-*k* filters, we find that the cutoff frequency of an *m*-derived filter may be expressed as

$$\omega_1 = \frac{2\,(R/L_1)}{\sqrt{4\dfrac{L_2}{L_1}+1}}. \tag{7}$$

To remove L_1 from the equation, note that the cutoff frequency may also be expressed as

$$\omega_1 = \frac{2m}{\sqrt{L_1 C}}, \tag{8}$$

while the characteristic impedance is given by

$$R = \sqrt{\frac{L_1}{C}}. \tag{9}$$

Combining these last three equations allows us to solve for L_2:

$$L_2 = \frac{(1-m^2)R}{2m\omega_1}. \tag{10}$$

Solving Eqn. 8 and Eqn. 9 for L_1 and C yields

$$C = \left(\frac{2m}{\omega_1}\right)\frac{1}{R} \tag{11}$$

and

$$L_1 = \left(\frac{2m}{\omega_1}\right)R. \tag{12}$$

Use of the foregoing equations requires that the designer have an idea of what value of *m* is desirable. As *m* approaches unity the response exhibits a monotonic rolloff (and therefore an increasing passband error), while passband peaking increases as *m* approaches zero. In practice a rather narrow range is encountered (say, within 25-30% of 0.5) as a compromise between these two behaviors, and the parameter *m* is commonly chosen equal to 0.6. This value yields a reasonably broad frequency range over which the transmission magnitude remains roughly constant. The following table enumerates (to more significant digits than are practically significant) some of the more relevant characteristics of *m*-derived filters, for the specific value of $m = 0.6$. As with Table 1, the parameter *n* is the number of complete T- (or π-) sections used in the filter. The column labeled "minimum stopband attenuation" gives the worst-case value of attenuation above the transmission notch frequency, where the filter response pops back up.

TABLE 2. Characteristics of ideal *m*-derived filters ($m = 0.6$)

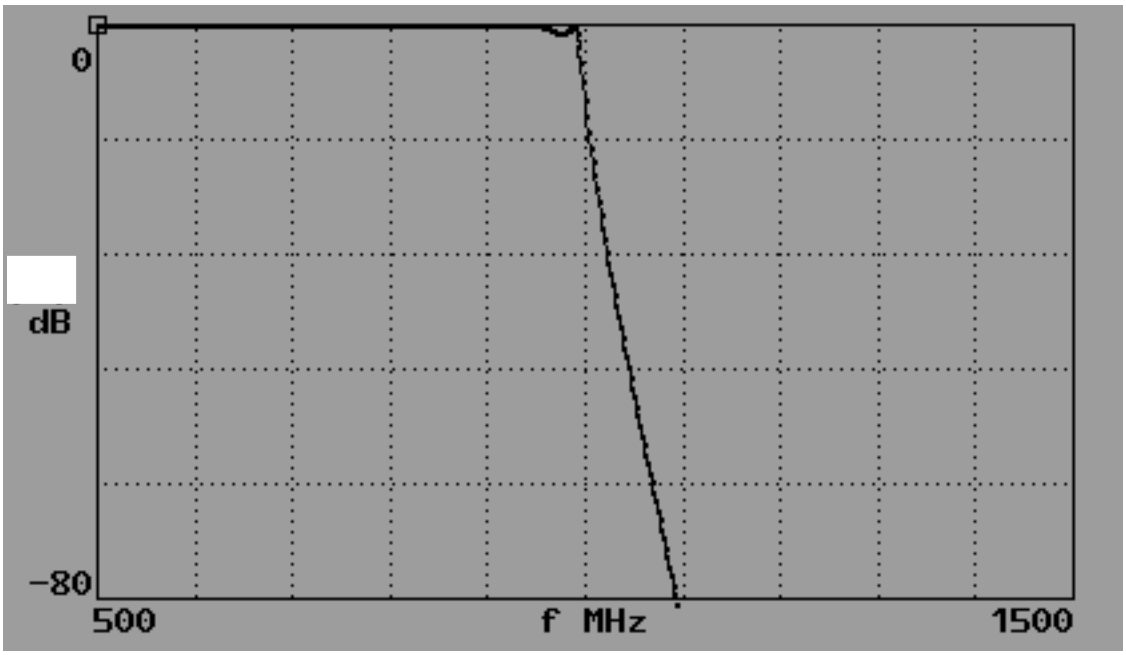
n	***Attenuation at cutoff frequency (dB)***	***Normalized –3dB Bandwidth***	***Normalized –6dB Bandwidth***	***Normalized –10dB S_{11} Bandwidth***	***Minimum stopband attenuation (dB)***
0	1.34	1.031	1.063	0.965	8.21
1	3.87	0.993	1.013	0.956	21.24
2	6.27	0.988	0.999	0.969	34.25
3	8.30	0.989	0.996	0.979	47.09
4	10.00	0.991	0.995	0.954	59.81
5	11.44	0.993	0.996	0.961	72.43

Note that the cutoff frequency and –3dB bandwidth are much more nearly equal than for the prototype constant-*k* case (the worst-case difference here is about 3%). The bandwidth

over which the return loss exceeds 10dB is also a much greater fraction of the cutoff frequency (above 95%, in fact). Note also that the minimum stopband attenuation increases by about 12-13dB per increment in *n* for this range of values.

The following figure illustrates how the use of *m*-derived sections can improve the magnitude response (note that the vertical axis now spans 80dB, rather than 50dB):

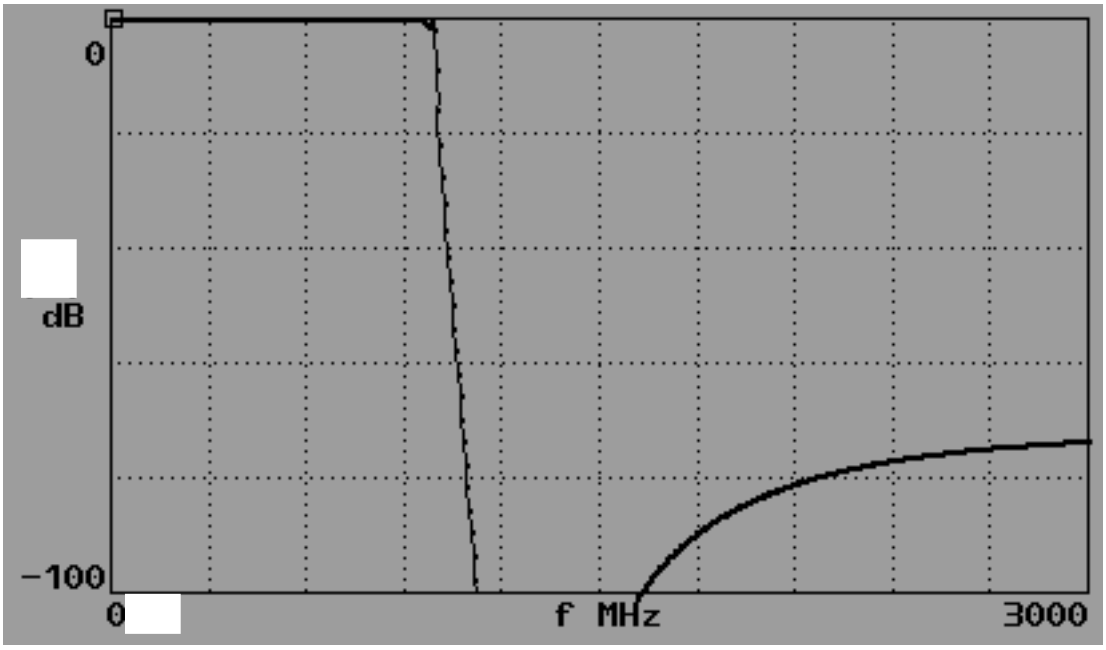
FIGURE 6. Frequency response of six-section *m*-derived low pass filter (*m* = 0.6, *n* = 5)



Compared with Figure 4, this response shows significantly less passband ripple, as well as a much faster transition to stopband, owing to the stopband notch.

On the frequency scale of Figure 6, the characteristic notch is invisible, as is the popping-up of the response at higher frequencies. The following plot shows these features more clearly:

FIGURE 7. Frequency response of *m*-derived low pass filter, plotted over wider range



Aside from the potential for improved flatness over the passband, the notches that are inherent in *m*-derived filters can be used to null out interfering signals at a specific frequency (or frequencies, if sections with differing values of *m* are used). Later, we will see that judiciously distributed notches can be combined with passband ripple to produce what are known as elliptic or Cauer filters whose responses exhibit even more dramatic transitions from passband to stopband.

If the precise location of a notch is of importance, it is helpful to know that the frequency of the null ω_∞ is related to *m* as follows:

$$\frac{\omega_\infty}{\omega_1} = \frac{1}{\sqrt{1-m^2}}, \tag{13}$$

so that the value of *m* needed to produce a notch at a specified frequency ω_∞ is

$$m = \sqrt{1-\left(\frac{\omega_1}{\omega_\infty}\right)^2}. \tag{14}$$

A value of 0.6 for *m* corresponds to a notch frequency that is a factor of 1.25 times the cutoff frequency.

Table 3 summarizes the design of constant-*k* and *m*-derived low pass filters. Component values (again, to many more digits than are practically relevant) are for the specific case of a termination (and source) resistance of 50Ω and a cutoff frequency of 1GHz. The left two

columns are for the simple constant-*k* case, and the last three columns give values for the specific *m*-derived case where $m = 0.6$.

TABLE 3. Component values for 1GHz constant-*k* and *m*-derived filters ($Z = 50\Omega$ $m = 0.6$)

L	C	L_1	C	L_2
15.9155nH	6.3662pF	9.5493nH	3.8197pF	4.2441nH

For filters with a cutoff frequency other than 1GHz, simply multiply all component values by the ratio of 1GHz to the desired cutoff frequency. For a different characteristic impedance, multiply all component *impedances* by the ratio of the desired impedance to 50Ω.

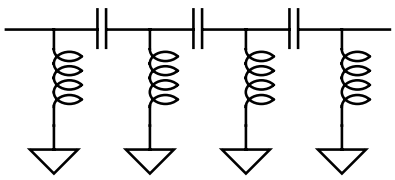
One may also combine ordinary constant-*k* and *m*-derived sections because the individual sections for both are constant-*k* in nature. Such a *composite* filter may be desirable, for example, to effect a compromise between flatness and the production of notches at specific frequencies. Unfortunately, design of such a filter is very much an ad hoc affair. One simply mixes and matches sections as seems sensible, then simulates to verify if the design indeed functions satisfactorily.

3.1.1 High-pass, bandpass and bandstop shapes

At least in principle, a high-pass constant-*k* filter is readily constructed from the low pass constant-*k* prototype simply by swapping the positions of the inductors and capacitors; the values remain the same. Thus one may design, say, a 1GHz constant-*k* low-pass filter using the values of Table 3, then interchange the *L*s and *C*s to synthesize a 1GHz high-pass filter.

The reason for the qualifier “at least in principle” is that high-pass filters typically exhibit serious deviations from desired behavior. These deviations motivate microwave filter designers to avoid high-pass filters based on lower frequency prototypes. Although there are many ways – too numerous to mention, in fact – in which a practical filter of any kind can fall short of expectations, perhaps the following lumped high pass filter example will suffice to illustrate the general nature of the problem. Specifically, consider:

FIGURE 8. High pass filter?



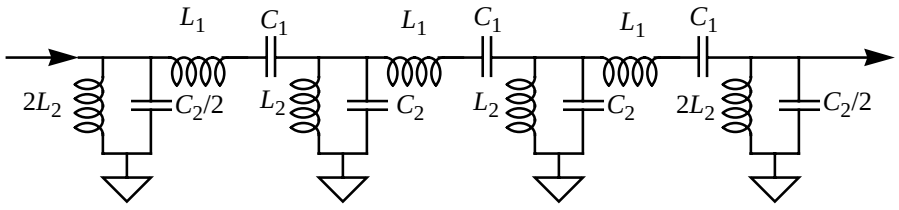
Every practical inductor is shunted by some capacitance, and thus exhibits a resonance of its own. Above the resonant frequency, the “inductor” actually appears as a capacitance. Similarly every practical capacitance has in series with it some inductance. Above the cor-

responding series resonance, the capacitor actually appears inductive. Hence, at sufficiently high frequencies, our high pass filter actually acts as a low pass filter.

A complementary effect afflicts low pass filters where, at high frequencies, it is possible for the response to pop back up.

As a practical workaround, it is traditional to employ a bandpass filter with a sufficiently wide passband to approximate the desired filter shape. Of course, that solution presupposes knowledge of how to construct bandpass filters. Fortunately, the constant-*k* structure works here, too (we'll later examine alternative bandpass implementations as well). As a general strategy for deriving a bandpass filter from a low-pass prototype, replace the inductance of a low-pass prototype with a series *LC* combination, and the capacitance with a parallel *LC* combination:

FIGURE 9. Bandpass constant-*k* filter example using two cascaded T-sections



Unlike our previous figures, the individual T-sections are not shown, in order to simplify the schematic.

Note that this structure continues to exhibit the correct qualitative behavior even if inductors ultimately become capacitors and vice-versa. This property is fundamental to the potentially reduced sensitivity of this topology to parasitic effects.

The formula for the inductance L_1 of the series resonator is the same as that for the inductance in the prototype low pass filter, except that the *bandwidth* (defined by the difference between the upper and lower cutoff frequencies) replaces the cutoff frequency. The capacitance C_1 is then chosen to produce a series resonance at the center frequency (defined here as the geometric mean of the two cutoff frequencies[12]). Hence

$$L_1 = \frac{2}{(\omega_2 - \omega_1)} R \tag{15}$$

and

12. In some of the literature, it is unfortunately left unclear as to what sort of mean should be used. For the common case of small fractional bandwidths, this ambiguity is acceptable, for there is then little difference between an arithmetic and geometric mean. Practical component tolerances make insignificant such minor differences. However, the discrepancy grows with the fractional bandwidth, and the error can become quite noticeable at large fractional bandwidths if the arithmetic mean is used.

$$C_1 = \frac{(\omega_2 - \omega_1)}{2\omega_0^2}\frac{1}{R}. \tag{16}$$

Similarly, the equation for the capacitance of the low-pass prototype is modified for the bandpass case by replacing the cutoff frequency with the bandwidth. The resonating inductance is again chosen to produce a resonance at the center frequency:

$$C_2 = \frac{2}{(\omega_2 - \omega_1)}\frac{1}{R} \tag{17}$$

and

$$L_2 = \frac{(\omega_2 - \omega_1)}{2\omega_0^2}R. \tag{18}$$

Values for a constant-*k* bandpass filter with cutoff frequencies of 950MHz and 1.05GHz (corresponding to a center frequency of approximately 998.75MHz) are given in the following table:

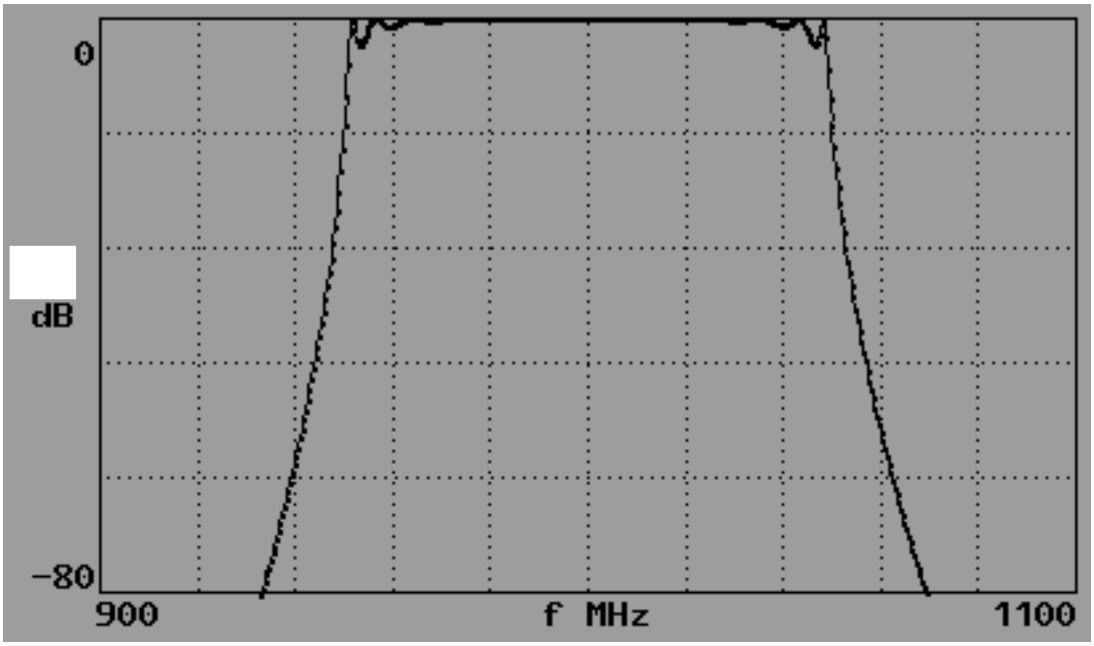
TABLE 4. Component values for a 100MHz bandwidth, constant-*k* bandpass filter at 1GHz

L_1	C_1	L_2	C_2
159.15nH	0.15955pF	0.39888nH	63.662pF

As is its low-pass counterpart, the bandpass filter is terminated in half-sections. Each half-section consists of components of value $L_1/2$, $2C_1$, $2L_2$ and $C_2/2$. The resulting filters have the same characteristics enumerated in Table 1, where the bandwidth normalizations continue to be performed, sensibly enough, with respect to the bandwidth, rather than the center frequency.

The following figure shows the frequency response of a bandpass filter derived from a low-pass constant-*k* filter with six sections ($n = 5$). The design bandwidth is 100MHz, centered at 1GHz. Not surprisingly, the behavior at the passband edges resembles that of the low-pass prototype.

FIGURE 10. Frequency response for bandpass filter derived from six-section constant-*k* (*n* = 5)



For a different bandwidth, multiply C_1 and L_2 by the ratio of the new bandwidth to 100MHz, and reduce L_1 and C_2 by the same factor. For a different center frequency, reduce C_1 and L_2 each by the square of the ratio of the new center frequency to 1GHz. Finally, for a different characteristic impedance, increase the impedance of all four components by the ratio of the new impedance to 50Ω.

The bandpass filter can be converted into a bandstop (also known as a band-reject) filter simply by swapping the positions of the series and parallel resonators. As in the conversion from low pass to high pass, the values remain unchanged.

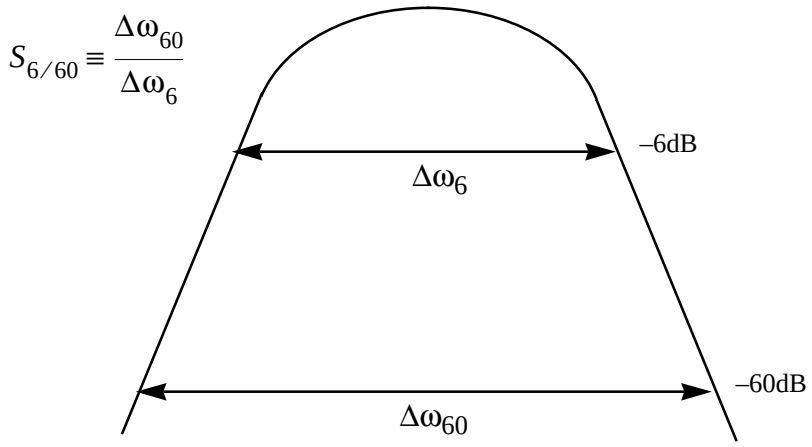
From the tables and examples given, it is clear that the constant-*k* and *m*-derived filters are extremely simple to design, since they consist of identical iterated sections (plus a terminating half-section on each end). This simplicity is precisely their greatest attribute. In exchange for this ease of design, however, the foregoing procedures neglect certain details (such as passband ripple) because they do not incorporate any specific constraints on response shape. It is clear from the tables, for example, that the cutoff frequency doesn't correspond to a certain fixed attenuation value, such as –6dB, and monotonicity is far from guaranteed. Stopband behavior is similarly mysterious. Finally, because the source and load terminations are assumed equal in value, any necessary impedance transformations have to be provided separately. It should seem reasonable, however, to expect that a more advanced synthesis technique might, on occasion, accommodate impedance transformation as a natural accompaniment to the filtering operation. Shortcomings such as these explain why there are alternative filter design approaches. Because the relative merits of these alternatives are best appreciated after identification and definition of key filter performance metrics, we now consider a brief sidebar and introduce these parameters.

4.0 Filter Classifications and Specifications

Filters may be classified broadly by their general response shapes – e.g., low-pass, band-pass, band reject and high-pass – and further subdivided according to bandwidth, *shape factor* (or skirt selectivity), and amount of ripple (in either the phase or magnitude response, and in either the passband, stopband, or both). This subdivision is an acknowledgment that ideal, brickwall filter shapes are simply unrealizable (not merely impractical). Different approaches to approximating ideal characteristics result in different tradeoffs, and the consequences of these compromises require characterization.

Bandwidth is perhaps the most basic descriptive parameter, and is conventionally defined using –3dB points in the response. However, it is important to recognize that 3dB is quite an arbitrary choice (there is nothing fundamental about the half-power point, after all), and we will use other bandwidth definitions that may be more appropriate from time to time. It is certainly an incomplete specification, because there are infinitely many filter shapes that share a common –3dB bandwidth. *Shape factor* is an attempt to convey some information about the filter's response at frequencies well removed from the –3dB point. It is defined as the ratio of bandwidths measured at two different attenuation values (i.e., values at two different points on the *skirt*). As an arbitrary example a "6/60" shape factor specification is defined as the bandwidth at –60dB attenuation, divided by the bandwidth at –6dB attenuation:

FIGURE 11. Illustration of 6/60 shape factor



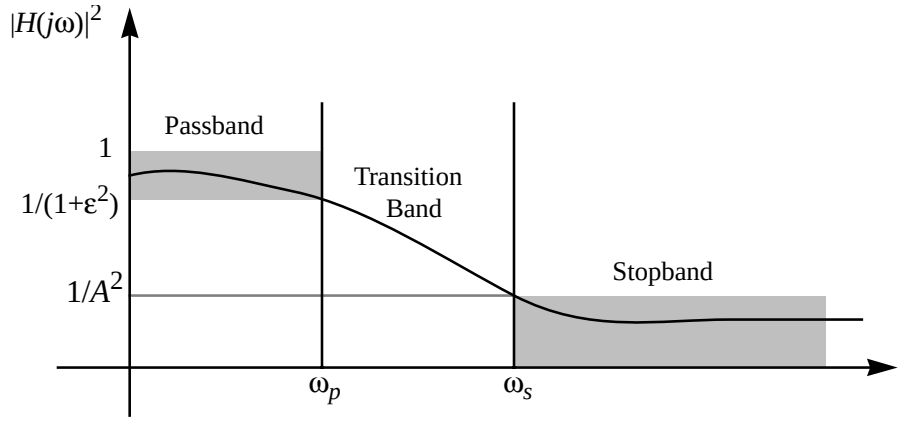
Clearly from the definition of shape factor, values approaching unity imply response shapes that approach infinitely steep transitions from passband to stopband. A single-pole lowpass filter (or a standard single-*LC* bandpass resonator) has a 6/60 shape factor of roughly 600, a value generally regarded as pathetically large.[13] This trio of numbers is easily remembered, though, because of the decimal progression.

13. The actual number is closer to 577, but has less of a mnemonic value than 600.

Because the relevance of a given shape factor depends very much on context, there cannot be a single, universally relevant definition. Thus although 6/60 happens to be a common one, other specifications are often encountered.

As stated earlier, the inability of practical filters to provide perfectly flat passbands and infinitely steep transitions to infinitely attenuating stopbands implies that we must always accept approximations to the ideal. In the best case we have the opportunity to quantify and specify bounds on the approximation error. The traditional way of doing so is to specify the following parameters:

FIGURE 12. General filter response template (shown for the low pass case)



Note that the square of the magnitude is plotted in the figure, rather than the magnitude itself, because it is proportional to power gain. This convention isn't universally followed, but it is quite common because of the RF engineer's typical preoccupation with power gain.

Note also the pervasiveness of reciprocal quantities on the vertical axis. This annoying feature is avoided by plotting attenuation, rather than gain, as a function of frequency, explaining why many treatments present data in precisely that manner.

Note further that the filter response template accommodates some amount of variation within the passband (whose upper limit is denoted ω_p), with a maximum permitted deviation of $1/(1+\varepsilon^2)$. Additionally, a finite transition between the passband and stopband (whose lower frequency limit is denoted ω_s) is also permitted, with a minimum allowed power attenuation of A^2 in the stopband. Specification of these parameters thus allows the design of real filters. We now consider several important classes of approximations which make use of these parameters.

5.0 Modern Filters: Common Approximations

The constant-*k* filter's limitations ultimately derive from a synthesis procedure which ignores the control over filter response afforded by direct manipulation of the pole (and zero) locations. This limitation is a natural consequence of the transmission-line theoretical basis for constant-*k* filters; because transmission lines are infinite-order systems, con-

sideration of pole locations there would be unnatural, and in any case would lead to numerous analytical difficulties.

However if one no longer insists on treating filters from a transmission line viewpoint, these difficulties disappear (but are replaced by new ones). Additional, and highly powerful, techniques then may be brought to bear on the filter analysis and synthesis problem. In this section, we underscore this point by following a synthesis procedure not possible with the constant-*k* filter: starting from a specification of a desired frequency response, compute a corresponding pole-zero constellation, and then synthesize a lumped network that exhibits the prescribed characteristics.

5.1 Butterworth filters

Some applications are entirely intolerant of ripple, limiting the number of options for response shape. As do all practical filters, the Butterworth seeks to approximate the ideal rectangular brickwall shape. The Butterworth filter's monotonic response shape minimizes the approximation error in the vicinity of zero frequency by maximizing the number of derivatives whose value is zero there. For a filter of order n, that maximum number happens to be $2n - 1$. As the filter order approaches infinity, the filter shape progressively approximates better the ideal brickwall shape.

A natural, but potentially undesirable consequence of a design philosophy which places greater importance on the approximation error at low frequencies is that the error grows as the cutoff frequency is approached. If this characteristic is indeed undesirable, one must seek shapes other than the Butterworth. Some of these alternatives are discussed in subsequent sections.[14]

The Butterworth's response magnitude (squared) as a function of frequency is given for the low pass case by the following expression:

$$|H(j\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}}, \tag{19}$$

where ω_c is the frequency at which the power gain has dropped to 0.5.[15] The parameter n is the order of the filter, and equals the number of independent energy storage elements, as well as the power of ω with which the response magnitude ultimately rolls off. From the equation, it is straightforward to conclude that the response is indeed monotonic.

14. As will be discussed later, one of these alternatives, the Type II Chebyshev, actually achieves better passband flatness than the Butterworth (making it flatter than maximally flat), by permitting stopband ripple, while preserving a monotonic passband response.

15. Although not rigorously correct (because of the possibility of unequal input and output impedances), we will frequently use the term "power gain" interchangeably with the more cumbersome "response magnitude squared."

In designing a Butterworth filter, one may specify ω_c directly, but to maintain consistency with the template of Figure 12, let us *derive* ω_c from the other filter parameters. First we may express the power gain at the passband and stopband edges as follows:

$$\frac{1}{1+\varepsilon^2} = \frac{1}{1+\left(\frac{\omega_p}{\omega_c}\right)^{2n}} \tag{20}$$

and

$$\frac{1}{A^2} = \frac{1}{1+\left(\frac{\omega_s}{\omega_c}\right)^{2n}}. \tag{21}$$

Solving these two equations for the required filter order, *n*, yields

$$n = \frac{\ln\left(\varepsilon/\sqrt{A^2-1}\right)}{\ln\left(\frac{\omega_p}{\omega_s}\right)}. \tag{22}$$

Thus, once the attenuation at the passband edge, minimum attenuation at the stopband edge, and the frequencies of those edges are specified, the required filter order is immediately determined. Because Eqn. 22 generally yields non-integer values, one must choose the next higher integer as the filter order. In that case, the resulting filter will exhibit characteristics that are superior to those originally sought. One way to use the "surplus" performance is to retain the original ω_p, in which case the filter will exhibit greater attenuation at ω_s than required. Alternatively, one may instead retain the original ω_s, in which case the filter exhibits smaller attenuation (i.e. smaller error) at the passband edge than originally targeted. Or, one may elect a strategy that is intermediate between these two choices.

Pursuing the strategy of retaining the originally sought performance at the passband edge, the –3dB corner frequency ω_c may be computed from the foregoing equations as

$$\omega_c = \frac{\omega_p}{\varepsilon^{1/n}}. \tag{23}$$

Alternatively, Eqn. 21 may be solved for a (generally different) ω_c derived from a specification of ω_s.

Because its approximation error is very small near DC, the Butterworth shape is also described as maximally flat.[16] However it is important to recognize that *maximally* flat does not imply *perfectly* flat.[17] Rather, it implies the flattest passband that can be

achieved, subject to the constraint of monotonicity (later, we will see that it is possible to have an even flatter passband response if we are willing to permit ripple in the stopband).

As a design example, let us continue the exercise that we began with the constant-*k* topology. We now have the ability to specify more filter parameters than in that case, so we'll do so. Here, arbitrarily allow a 1dB loss (gain of 0.794) at the passband edge of 1GHz, and require a 30dB factor of minimum attenuation at a 3GHz stopband edge.

From the passband specification, we find that ε is approximately 0.5088. From the stopband specification, we see that A^2 is 1000. As a result, the minimum filter order required to meet the specifications is

$$n = \frac{\ln(0.5088/\sqrt{999})}{\ln\left(\frac{1}{3}\right)} \approx 3.76, \tag{24}$$

which we round upward to four. Choosing to meet precisely the specification at the passband edge, we find that the corresponding value of ω_c is approximately

$$\omega_c = \frac{\omega_p}{\varepsilon^{1/n}} \approx 7.44 Grps. \tag{25}$$

From this point, we would typically consult a table of component values for a fourth-order filter, scaling the values for the desired cutoff frequency (and, possibly, impedance level).[18] As a final check, it is always wise to simulate the proposed filter, just to make sure that no computational errors (or typographical errors – some published tables have incorrect entries!) have corrupted the design. In demanding applications, simulation is also valuable for assessing the sensitivities of the filter to practical variations in component values, or to other imperfections (such as finite *Q* or parasitics).

5.2 Chebyshev (equiripple or minimax) filters

Although monotonicity certainly has an esthetic appeal, insisting on it constrains other valuable filter shape properties. These include the steepness of transitions from passband to stopband, as well as the stopband attenuation for a given filter order. Alternative filters, based on non-monotonic frequency response, are named after the folks who invented them, or who developed the underlying mathematics. The Chebyshev filter, an example of

16. This term was evidently introduced by V. D. Landon in his paper, "Cascade Amplifiers with Maximal Flatness," *RCA Review*, vol. 5, pp. 347-362, January 1941. Coining of the term thus follows by more than a decade Butterworth's own exposition of the subject in "On the Theory of Filter Amplifiers," *Wireless Engr.*, vol. 7, pp. 536-541, Oct. 1930. Although others published similar results earlier, *Butterworth* and *maximal flatness* are now seemingly linked forever.

17. In this way, "maximally flat" is used a bit like "creme filling" in describing the ingredients of an Oreo™ cookie; it means something a little different from how it initially sounds.

18. Later, we will present a synthesis method that allows computation of component values directly.

the latter, allows a reduction in filter order precisely by relaxing the constraint of monotonicity.[19] In contrast with the Butterworth approximation, which is preoccupied with minimizing error at low frequencies, the Chebyshev minimizes the maximum approximation error (relative to the ideal brickwall shape) throughout the entire passband. The resulting *minimax* response shape thus exhibits some ripple, the amount of which may be specified by the designer. For a given order, the Chebyshev filter shape offers a more dramatic transition from passband to stopband than a Butterworth offers. The steepness of the transition is also a function of the passband ripple one allows; the greater the permissible ripple, the steeper the transition.

A consequence of minimizing the maximum error is that the ripples of a Chebyshev response are all of equal amplitude. A rigorous proof of the minimax optimality of an equiripple shape is surprisingly involved, so we won't attempt one here. However, it should seem intuitively reasonable that equiripple behavior would be optimal in the minimax sense for, if any one error peak were larger than the others, a better approximation could probably be produced by reducing it, at the cost of increasing the size of one or more of the others. Such tradings-off would proceed until all error peaks were equal, because nothing would then be left to trade for anything else.

Similar advantages also accrue if the stopband, rather than the passband, is allowed to exhibit ripple. The inverse Chebyshev filter (also known as a Type II Chebyshev filter) is based on this idea, and actually combines a flatter-than-Butterworth passband with an equiripple stopband.

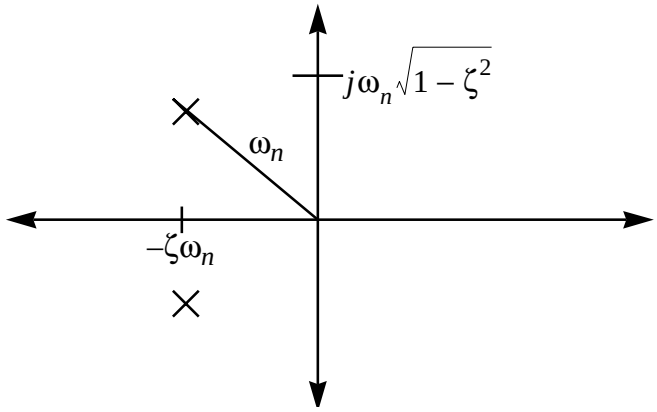
To understand how the simple act of allowing either stopband or passband ripple provides these advantages, we need to review the properties of a complex pole pair. First recall that one standard (and perfectly general) form for the transfer function of such a pair is:

$$H(s) = \frac{1}{\frac{s^2}{\omega_n^2} + \frac{2\zeta s}{\omega_n} + 1}, \tag{26}$$

where ω_n is the distance to the poles from the origin, and ζ (zeta) is the damping ratio:

19. Pafnuti L'vovich Chebyshev (1821-1894) did no work on filters at all. In fact he developed his equations during a study of mechanical linkages used in steam engines (see his posthumously published "Théorie des mécanismes connus sous le nom de parallélogrammes," (Theory of mechanisms known under the name of parallelograms)), *Oeuvres*, vol. I, St. Petersburg, 1899. "Parallelograms" translate rotary motion into an approximation of rectilinear motion. By the way, the spelling of his name here is just one of many possible transliterations of Pafnutiy L¡voviq Qeb[wev.

FIGURE 13. Two-pole constellation

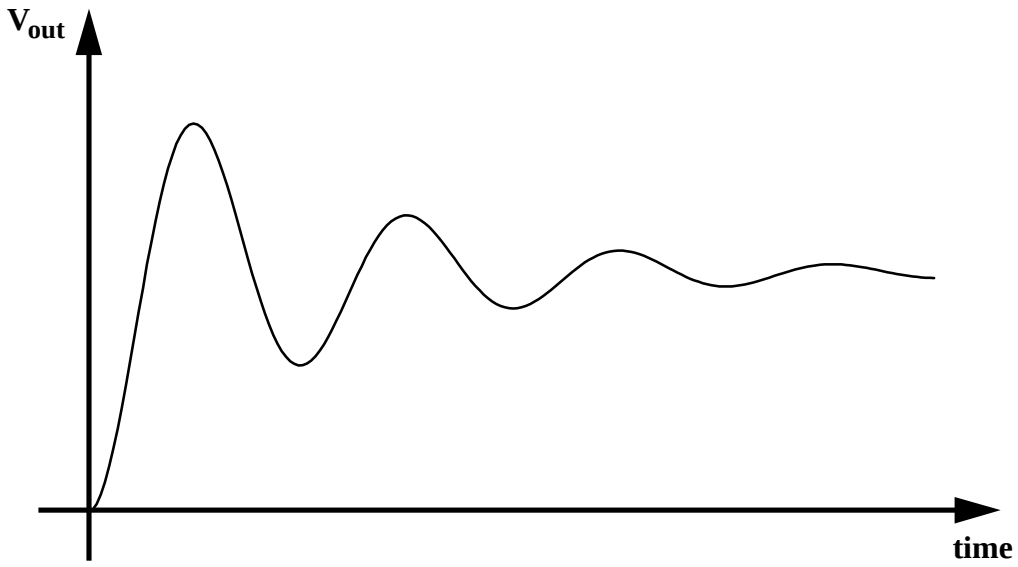


A zero damping ratio corresponds to purely imaginary poles, and a damping ratio of unity corresponds to a pair of poles coincident on the real axis. The former condition applies to an oscillator, and the latter defines critical damping. Above a damping ratio of unity the two poles split, with one moving toward the origin, the other toward minus infinity, all the while remaining on the real axis. Whatever the value of damping, the frequency ω_n always equals the geometric mean of the pole frequencies.

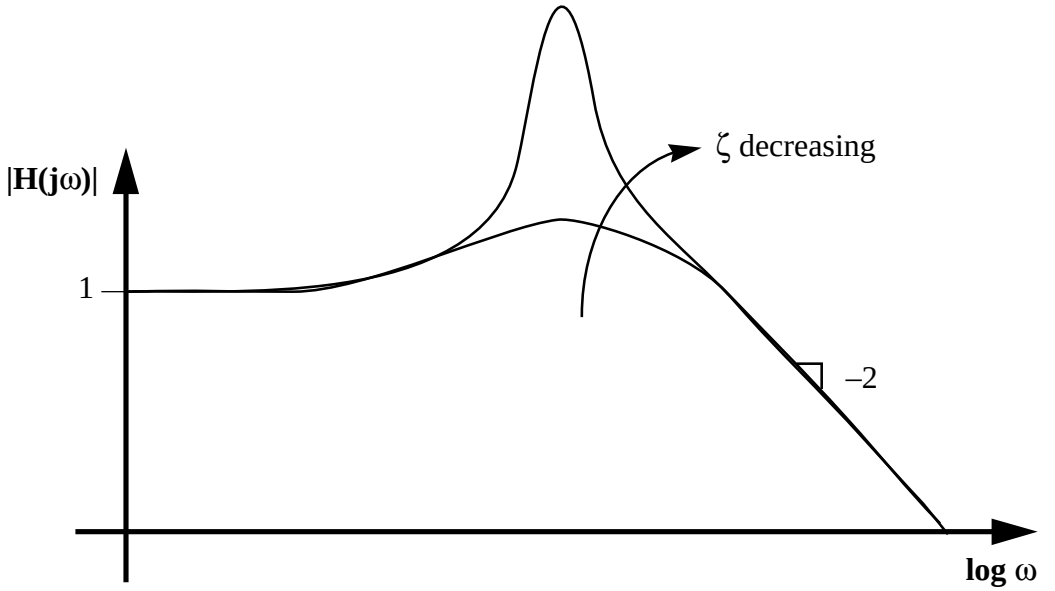
The property that is most relevant to the subject of filter design is the dependency of the frequency response shape on the damping ratio. While it is true that all zero-free two pole systems have a frequency response that ultimately rolls off as ω^{-2}, the frequency response magnitude, and the slope, *in the vicinity of the peak* are very much functions of the damping ratio, increasing as ζ decreases (Figure 14b). For damping ratios above $1/\sqrt{2}$, the frequency response exhibits no peaking. Below that value of ζ, peaking increases without bound as the damping ratio approaches zero. For small values of ζ, the peak gain is inversely proportional to damping ratio. Stated alternatively, lower damping ratios lead to greater ultimate attenuation, relative to the peak gain, and to slopes that are normally associated with higher (and perhaps much higher) order systems.

Now consider ways a filter might exploit this ζ-dependent behavior. Specifically, suppose we use a second-order section to improve the magnitude characteristics of a single-pole filter. If we arrange for the peak of the second-order response to compensate (boost) the response of the first-order section beyond where the latter has begun a significant rolloff, the frequency range over which the magnitude of the cascade remains roughly constant can be increased. At the same time, the rolloff beyond the compensation point can exhibit a rather high initial slope, providing an improved transition from passband to stopband. Clearly, additional sections may be used to effect even larger improvements, with each added section possessing progressively smaller damping ratios. This latter requirement stems from the need to provide larger boosts to compensate for ever larger attenuations.

FIGURE 14. Step- and frequency-response of second-order lowpass transfer characteristic



a) Step response

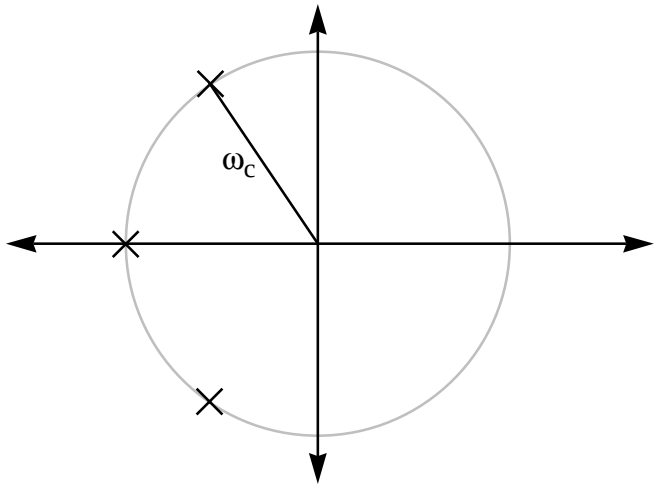


b) Frequency response

Having developed this understanding, we may revisit the Butterworth and Chebyshev approximations. The Butterworth condition results when the poles of the transfer characteristic are arranged so that the modest amount of frequency response peaking of a complex pole pair offsets, to a certain extent, the rolloff of any real pole present. The resulting combination exhibits roughly flat transmission magnitude over a broader frequency range than that of either the real pole or complex pair alone. The result is that all of the poles lie on a semicircle in the *s*-plane, distributed as if there were twice as many poles disposed at equal angles along the circumference, the right half-plane poles being ignored.[20] A third-

order Butterworth, for example, has a single pole on the real axis, and a complex conjugate pair at 60° angles with the real axis. The distance from the origin to the poles is the 3dB cutoff frequency.

FIGURE 15. Pole constellation for third order Butterworth low pass filter



The element values, normalized to a 1Ω impedance level, and to a 1 rad/sec passband edge, for an *n*th-order Butterworth low-pass filter are given by the following set of equations:

$$g_0 = 1 \tag{27}$$

and

$$g_k = 2\sin\left[\frac{(2^k - 1)\,\pi}{2^n}\right], \tag{28}$$

where *k* ranges from 1 to *n*, and

$$g_{n+1} = 1. \tag{29}$$

Conversion into a bandpass filter is easily achieved using the same transformations used in the constant-*k* case.

The Chebyshev filter goes further by allowing passband (or stopband) ripple. Continuing with our third-order example, the response of the real pole is allowed to drop below the low-frequency value by some specified amount (the permissible ripple) before the complex pair's peaking is permitted to bring the response back up. The damping ratio of the complex pair must be lower than that in the Butterworth case to produce enough additional peaking to compensate for the greater attenuation of the real pole. A side effect of

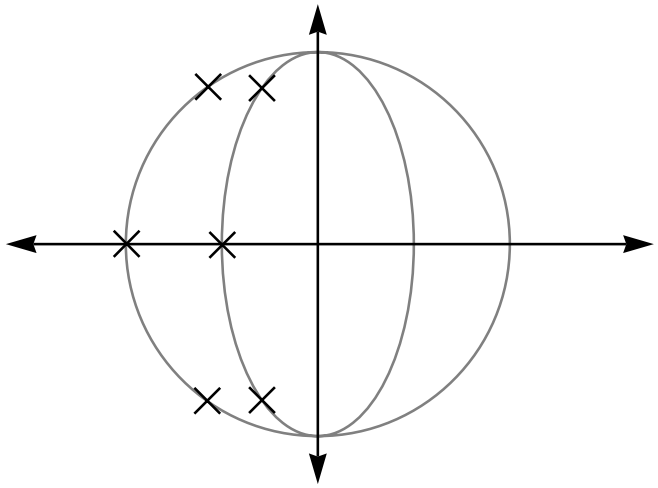
20. Okay, perhaps it isn't quite "intuitively obvious," it is true, but finding the roots of Eqn. 19 to discover the factoid about Butterworth poles lying on a circle isn't all that bad.

this lower damping is that there is a more dramatic rolloff beyond the cutoff frequency. In this manner the Chebyshev filter permits the designer to trade passband flatness for better stopband attenuation.

Although it is even less intuitively obvious, the poles of a Chebyshev low pass filter are located along a (semi)ellipse, remarkably with imaginary parts that are equal to those of a corresponding Butterworth low pass filter.[21] Increasing the eccentricity of the ellipse increases the ripple.

Inverse Chebyshev filters have poles located at the reciprocals of the "normal" Chebyshev, and purely imaginary zeros distributed in some complicated fashion. The resulting pole-zero constellation roughly resembles the Greek letter Ω rotated counter-clockwise by 90°.

FIGURE 16. Third order Butterworth and Chebyshev low pass filter pole constellations



Mathematically, the Chebyshev response is of the general form

$$|H(j\omega)|^2 = \frac{1}{1+\varepsilon^2 C_n^2\left(\frac{\omega}{\omega_p}\right)}, \tag{30}$$

where ω_p once again is the frequency at which the response magnitude squared has dropped to a value

$$\frac{1}{1+\varepsilon^2}, \tag{31}$$

as in the Butterworth case. For self-evident reasons ε is known as the ripple parameter, and is specified by the designer. The function $C_n(x)$ is known as a Chebyshev polynomial of

21. There are many references that provide excellent derivations of the Butterworth and Chebyshev conditions. A particularly enlightening derivation may be found in chapters 12 and 13 of R. W. Hamming's volume, *Digital Filters*, Prentice-Hall, 2nd ed., 1983.

order *n*. The most relevant property of such polynomials is that they oscillate between –1 and +1 as the argument *x* varies over the same interval. Outside of this interval the magnitude grows rapidly (as x^n in fact). The filter's (power) response thus varies between 1 and $1/(1+\varepsilon^2)$ as the frequency increases from DC to ω_p. That entire frequency interval is often called the ripple passband, and the parameter ω_p the ripple bandwidth (or ripple cutoff frequency). In general the ripple passband differs from the more conventional –3dB bandwidth.

There are a couple of ways of generating Chebyshev polynomials algorithmically. One is through a recursion formula,

$$C_n(x) = 2xC_{n-1}(x) - C_{n-2}(x), \tag{32}$$

where $C_0 = 1$ and $C_1 = x$ (just to get you started). As can be seen from the recursion formula, the leading coefficient of Chebyshev polynomials is 2^{n-1}, a fact we shall use later in comparing Chebyshev and Butterworth polynomials.

Another method for generating the Chebyshev polynomials is in terms of some trigonometric functions, from which the oscillation between –1 and +1 (for $|x| < 1$) is directly deduced:

$$C_n(x) = \cos(n\cos^{-1}x) \text{ for } |x| < 1. \tag{33}$$

For arguments larger than unity, the formula changes a little bit:

$$C_n(x) = \cosh(n\cosh^{-1}x) \text{ for } |x| > 1. \tag{34}$$

Although it is probably far from obvious at this point, these functions are likely familiar to you as Lissajous figures, formed and displayed when sinewaves drive both the vertical and horizontal deflection plates of an oscilloscope. That is, suppose that the horizontal deflection plates are driven by a signal

$$x = \cos t, \tag{35}$$

so that

$$t = \cos^{-1}x. \tag{36}$$

Further suppose that the vertical plates are simultaneously driven by a signal

$$y = \cos nt. \tag{37}$$

Substituting Eqn. 36 into Eqn. 37 to remove the time parameter yields

$$y = \cos(n\cos^{-1}x), \tag{38}$$

which is seen to be the same as Eqn. 33. That is, what's displayed on an oscilloscope driven in this fashion is in fact the Chebyshev polynomial for that order *n*, for values of |*x*| up to one. Over that interval the function displayed looks very much like a sinusoid sketched on a piece of paper, wrapped around a cylinder, and then viewed from a distance.

A few Chebyshev polynomials are sketched crudely in the following figure, and expressions for the first ten Chebyshev polynomials are listed in Table 5:

FIGURE 17. Rough sketches of some Chebyshev polynomials

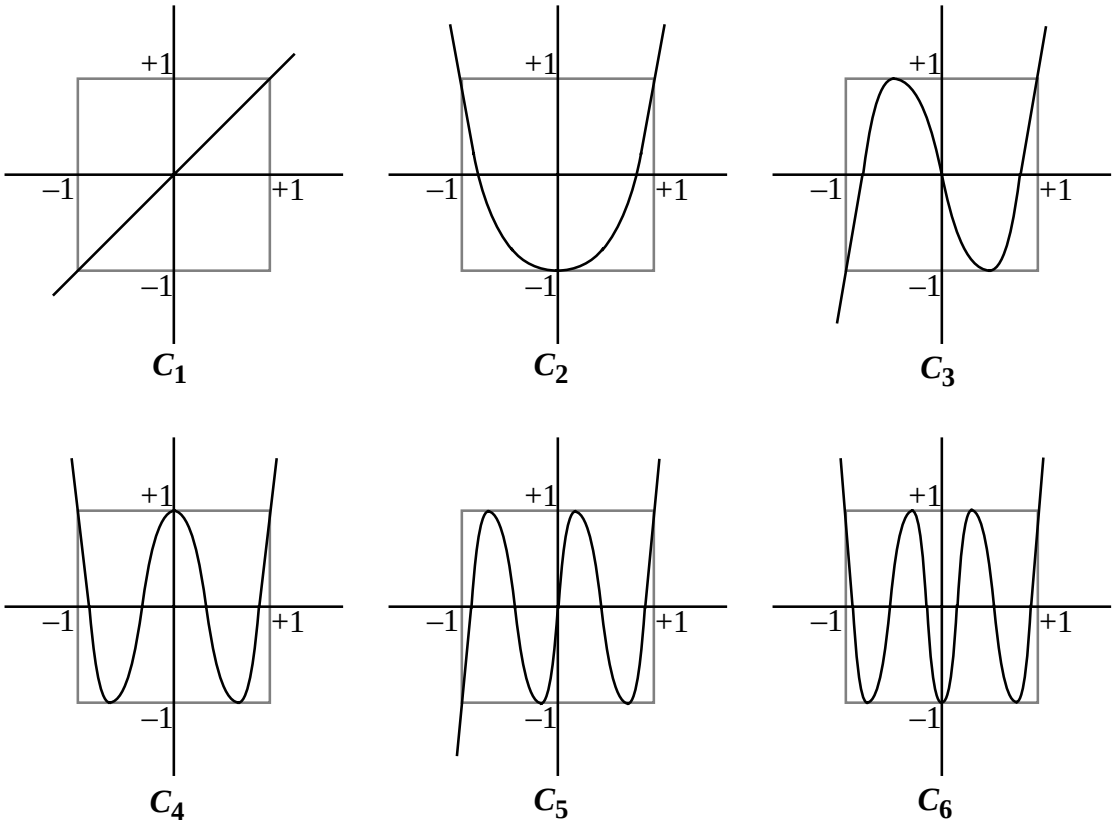


TABLE 5. First ten Chebyshev polynomials

Order, *n*	Polynomial
0	1
1	x
2	$2x^2 - 1$
3	$4x^3 - 3x$
4	$8x^4 - 8x^2 + 1$
5	$16x^5 - 20x^3 + 5x$
6	$32x^6 - 48x^4 + 18x^2 - 1$

TABLE 5. First ten Chebyshev polynomials

Order, *n*	Polynomial
7	$64x^7 - 112x^5 + 56x^3 - 7x$
8	$128x^8 - 256x^6 + 160x^4 - 32x^2 + 1$
9	$256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x$

From the foregoing equations, we may derive an expression for the filter order required to satisfy the specified constraints:

$$n = \frac{\cosh^{-1}\left(\sqrt{A^2 - \frac{1}{\varepsilon}}\right)}{\cosh^{-1}\left(\frac{\omega_s}{\omega_p}\right)}. \tag{39}$$

As with the Butterworth case, the order as computed by Eqn. 39 should be rounded upward to the next integer value. Again, the resulting "excess" performance can be used to improve some combination of passband and stopband characteristics.

One way in which the Chebyshev is superior to a Butterworth is in the ultimate stopband attenuation provided. At high frequencies, the Butterworth provides an attenuation that is approximately

$$\left|A\left(j\frac{\omega}{\omega_c}\right)\right|^2 \approx \left(\frac{\omega}{\omega_c}\right)^{2n}. \tag{40}$$

Compare that asymptotic behavior with that of a Chebyshev (with $\varepsilon = 1$ so that the –3dB frequency of the Butterworth corresponds to the passband edge of the Chebyshev):[22]

$$\left|A\left(j\frac{\omega}{\omega_c}\right)\right|^2 \approx 2^{2n-2}\left(\frac{\omega}{\omega_c}\right)^{2n}. \tag{41}$$

Clearly the Chebyshev filter offers higher ultimate attenuation, by an amount equal to 3dB($2n - 2$), for a given order. As a specific example, a 7th-order Chebyshev ultimately provides 36dB more stopband attenuation than does a 7th order Butterworth.

As another comparison, the relationship between the poles of a Butterworth and those of a Chebyshev of the same order can be put on a quantitative basis by normalizing the two filters to have precisely the same –3dB bandwidth. It also may be shown (but not by us) that the –3dB bandwidth of a Chebyshev may be reasonably well approximated by[23]

22. This comparison should not mislead you into thinking that such large ripple values are commonly used. In fact, Chebyshev filters are typically designed with ripple values below 1dB.

$$\cosh\left(\frac{1}{n}\sinh^{-1}\left(\frac{1}{\varepsilon}\right)\right). \tag{42}$$

Since the diameter of a Butterworth's circular pole constellation is the –3dB bandwidth, we normalize the Chebyshev's ellipse to have a major axis defined by Eqn. 42. The imaginary parts of the poles of a Chebyshev filter are the same as for the Butterworth, while the real parts of the Butterworth prototype are merely scaled by the factor

$$\tanh\left(\frac{1}{n}\sinh^{-1}\left(\frac{1}{\varepsilon}\right)\right) \tag{43}$$

to yield the real parts of the poles of a Chebyshev filter. Thus design of a Chebyshev filter is quite straightforward because it requires only a prototype Butterworth, and it's trivial to design the latter. Clearly, the Butterworth may be considered merely a special case of a Chebyshev, one for which the ripple parameter is zero.

There is one subtlety that requires discussion, however, and this concerns the source and termination impedances of a passive Chebyshev filter. From both the sketches and equations, it's clear that only odd-order Chebyshev polynomials have a zero value for zero arguments. Hence, the DC value of the filter transfer function will be unity for such polynomials (that is, the passband hits its ripple extremum at some frequency above DC). For even-order Chebyshev filters, however, the filter's transfer function starts off at a ripple extremum, with a DC power transmission value of $1/(1+\varepsilon^2)$, implying a termination resistance that is less than the source resistance. If, as is usually the case, such an impedance transformation is undesired, one must either use only odd-order Chebyshev filters, or add an impedance transformer to an even-order Chebyshev filter. As the former is less complex, it is the near universal choice to use only odd-order Chebyshev realizations in practice.

Finally, recognize that the elliptical pole distribution implies that the ratio of the imaginary to real parts of the poles, and hence the *Q*s of the poles, are higher for Chebyshevs than for Butterworths of the same order. As a result, Chebyshev filters are more strongly affected by the finite *Q* of practical components. The problem increases rapidly in severity as the order of the filter increases. This important practical issue must be kept in mind when choosing a filter type.

Element values for the Chebyshev filter are given by the following sequence of equations. First, compute four auxiliary quantities, whose significance may initially seem mysterious:[24]

$$\beta = ln\left(\coth\frac{L_{Ar}}{17.372}\right), \tag{44}$$

23. See, for example, M. E. Van Valkenburg, *Introduction to Modern Network Synthesis*, Wiley, 1960, pp. 380-381.

24. After close examination, these remain mysterious. Sorry. At least I can tell you that the 17.372 factor is $40\log_{10}e$, as if that helps.

where L_{Ar} is in dB;

$$\gamma = \sinh\frac{\beta}{2^n}; \tag{45}$$

$$a_k = \sin\left[\frac{(2^k-1)\,\pi}{2^n}\right]; \tag{46}$$

and

$$b_k = \gamma^2 + \sin^2\left(\frac{k\pi}{n}\right). \tag{47}$$

Once the values of the auxiliary parameters are known, the following equations yield the normalized element values:

$$g_0 = 1; \tag{48}$$

$$g_1 = \frac{2a_1}{\gamma}; \tag{49}$$

$$g_k = \frac{4a_{k-1}a_k}{b_{k-1}g_{k-1}}; \tag{50}$$

$$g_{n+1} = 1 \text{ for n odd}; \tag{51}$$

and

$$g_{n+1} = \coth^2\left(\frac{\beta}{4}\right) \text{for n even}. \tag{52}$$

5.3 Type II (Inverse) Chebyshev filters

We have alluded several times to the possibility of realizing a flatter-than-maximally flat transfer characteristic. The Type II (also known as an inverse or reciprocal) Chebyshev filter achieves such flatness by permitting ripple in the stopband, while continuing to insist on passband monotonicity.

The Type II filter derives from the Type I (ordinary) Chebyshev through a pair of simple transformations. In the first step, the Type I Chebyshev response is simply subtracted from unity, leading to the conversion of a low-pass filter into a high-pass one. Note that the resulting response is monotonic in the new passband. The second step replaces ω by $1/\omega$. Since high frequencies are thus mapped into low ones, and vice-versa, this second transformation converts the filter shape back into a low-pass response, but in a way that exchanges the ripple at low frequencies with ripple at high frequencies. This transforma-

tion thus restores a monotonic passband, and also happens to map the Type I passband edge into the stopband edge. Furthermore the larger the permissible stopband ripple, the flatter the passband response.

Mathematically, these transformations ultimately result in the following power response for a Type II filter:

$$|H(j\omega)|^2 = 1 - \frac{1}{1+\varepsilon^2 C_n^2\left(\frac{\omega_p}{\omega}\right)} = \frac{\varepsilon^2 C_n^2\left(\frac{\omega_p}{\omega}\right)}{1+\varepsilon^2 C_n^2\left(\frac{\omega_p}{\omega}\right)}. \tag{53}$$

Although the Type II filter is not encountered as often as the Butterworth, its relative rarity should not be taken to imply a corresponding lack of utility. Despite the superior flatness provided by the inverse Chebyshev, it appears that, for purely cultural reasons, the Butterworth filter continues to dominate in those applications where passband uniformity is allegedly prized.

5.4 Elliptic (Cauer) filters

Having seen that allowing ripple in the passband or stopband confers desirable attributes, perhaps it should not be surprising that the elliptic or Cauer filter further improves transition steepness by allowing ripple in both the passband and stopband simultaneously.[25] Just as a complex pole pair provides peaking, a complex zero pair provides notching. We've seen this behavior already, where the purely imaginary zeros of an *m*-derived filter provide notches of infinite depth. Cauer filters exploit this notching to create a dramatic transition from passband to stopband, at the expense of a stopband response that bounces back up beyond the notch frequency (again, just as in an *m*-derived filter, and for the same reasons). The name *elliptic* comes from the appearance of elliptic functions in the mathematics, and should not be confused with the elliptic pole distribution of a Chebyshev filter.

Elliptic filters have the following power transmission behavior:

$$|H(j\omega)|^2 = \frac{1}{1+\varepsilon^2 U_n^2\left(\frac{\omega}{\omega_c}\right)}, \tag{54}$$

where $U_n(x)$ is a *Jacobian elliptic function* of order *n*:[26]

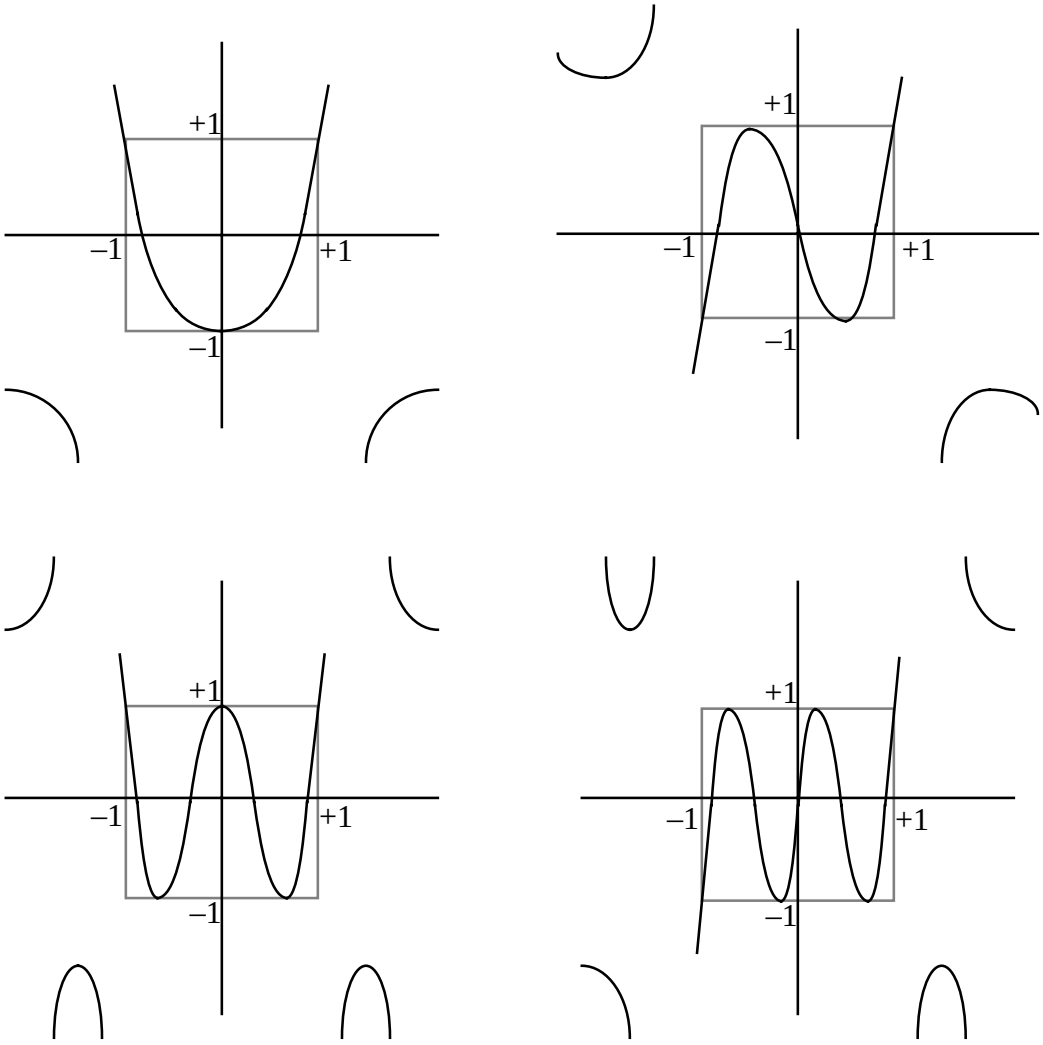
25. These are also sometimes known as Darlington or Zolotarev filters. Sidney Darlington (of "Darlington pair" fame in bipolar circuits) made major contributions in the field of network synthesis. Igor Ivanovich Zolotarev independently studied Chebyshev functions a decade or so before Chebyshev did.

26. These are named for the mathematician Karl Gustav Jacob Jacobi (1804-1851), who began studying these functions in the 1820s, at the start of his career.

$$U_n(x) = \int_0^x \frac{dy}{\sqrt{(1-y^2)\,(1-n^2y^2)}}. \tag{55}$$

These functions are messy enough that quantitative information about them is generally presented in tabular form (but not here, though; see, e.g., the oft-cited work by G.W. Spencely and R. M. Spencely, “Smithsonian Elliptic Function Tables,” Publication 3863, Smithsonian Institution, Washington, D. C., 1947). Suffice it to say that, just as Chebyshev polynomials do, these elliptic functions oscillate within narrow limits for arguments $|x|$ smaller than unity, and rapidly grow in magnitude for arguments outside of that range. However, unlike Chebyshev polynomials, whose magnitudes grow monotonically outside of that range, these elliptic functions oscillate in some fashion between infinity and a specified finite value. Hence the filter response exhibits stopband ripples, with a finite number of frequencies at which the filter transmission is (ideally) zero. The following figure shows crude sketches of the first several Jacobian elliptic functions, from which this behavior may be discerned:

FIGURE 18. Rough sketches of some Jacobian elliptic functions



The attenuation poles correspond to transmission zeros (notches), in the proximity of which the filter response changes rapidly. Thus, perhaps you can see how permitting such ripples in the stopband allows for a much more dramatic transition from passband to stopband, and thus allows one to combine the attributes of ordinary and inverse Chebyshev filters.

Wilhelm Cauer is the inventor whose deep physical insights (and intimate familiarity both with the notches of Zobel's *m*-derived filters, and with elliptic functions in general) allowed him first to recognize that this additional degree of freedom existed, and then to exploit it, even though he did not offer a formal mathematical proof of the correctness of his ideas.[27] At a time when minimizing component count was an obsession, Cauer was able to use fewer inductors than the best filters that were then in use. According to lore, publication of his patent reportedly sent Bell Labs engineers and mathematicians scurrying off to the New York City Public Library to bone up on the then-obscure (okay, still-obscure) literature on elliptic functions.[28]

6.0 Coupled Resonator Bandpass Filters

Up to now we've focused mainly on low pass filters, having derived other filter shapes from low pass prototypes. It is worthwhile to develop additional insights, however, so that we don't always have to return to the low pass case whenever we wish to design, say, a bandpass filter. This freedom, in turn, allows us to analyze and synthesize filter types that are not readily related to lumped networks at all.

We've seen that the poles, say, of a "good" filter aren't all coincident; they're distributed in some manner. Viewed from a broad perspective, then, the goal of filter design is to distribute the transfer function poles and zeros in some manner to achieve a desired response shape. This important idea is the basis for essentially all lumped filters, bandpass or otherwise. A particularly simple way to synthesize bandpass filters with a variety of response shapes is to exploit the *mode splitting* that occurs when two or more resonant systems interact. That is, when two identical resonators are connected together in some fashion, the poles of the resulting coupled system generally differ from those of the resonators in isolation. By controlling the degree of interaction (coupling) the pole locations can be adjusted to produce a desired response shape.

---

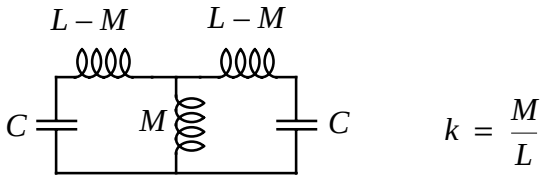
27. Cauer (1900-1945) became familiar with elliptic functions while studying at the University of Göttingen with the mathematician David Hilbert. Hilbert was as well known for his absentmindedness as for his mathematics. Once he suddenly asked a close friend, physicist James Franck, "Is your wife as mean as mine?" Though taken aback, Franck managed to respond, "Why, what has she done?" Hilbert answered, "I discovered today that my wife does not give me an egg for breakfast. Heaven only knows how long this has been going on."

It is unfortunate that stories about Cauer are not as lighthearted. Tragically, he was shot to death during the Soviet occupation of Berlin in the closing days of WWII, in a manner sadly reminiscent of the death of Archimedes (see http://www-ft.ee.tu-berlin.de/geschichte/th_nachr.htm).

28. M. E. Van Valkenburg, *Analog Filter Design*, Harcourt Brace Jovanovich, 1982, p. 379.

To illustrate this important idea, consider two simple *LC* resonators whose inductors are magnetically coupled to each other. Such a system may be modeled by representing the coupled inductors with a transformer. The transformer in turn is modeled as a T-connection of three inductors:

FIGURE 19. Coupled *LC* resonators



The inductance *L* is that which is present in each resonator in isolation. The mutual inductance *M* is a fraction of *L*, and depends on the magnitude of the coupling. The latter is captured in the coupling coefficient *k*, which ranges from zero to unity as the flux linkage of the magnetic fields of the two inductors increases from zero to 100%.

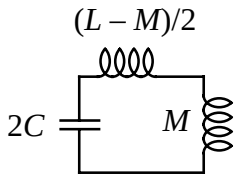
To find the resonant frequencies of the resulting 4th-order system[29] one can always employ a brute-force approach: Find the transfer function (first, one needs to define the input and output terminals), then solve for the roots of the denominator polynomial. This method is quite general, but also quite cumbersome, particularly for networks of order higher than two or three. Here, the network happens to be symmetrical, a situation that almost demands exploitation to simplify analysis by bypassing uninspired routes to the answer.

First recall what poles are. Yes, they are the roots of the denominator of the transfer function, but a deeper significance is that they are the natural frequencies of a network. That is, if the system is given some initial energy, the evolution of the system state *in the absence of any further input* takes place with characteristic frequencies whose values are those of the poles. Cleverly chosen initial conditions may excite only a subset of all possible modes at a time, thus converting a difficult high-order problem into a collection of more simply solved low-order ones. *Very* clever (or lucky) choices can even result in the excitation of a single mode at a time.

We may use this understanding to devise a simple method for finding the poles of our coupled resonator system. First, provide a common-mode excitation by depositing, say, an equal amount of initial charge on the two capacitors. Regardless of what the network does subsequently, we know by symmetry that the capacitor voltages must evolve the same way. Because the two capacitor voltages are thus always equal, we may short the capacitors together with impunity, resulting in the following network:

29. Despite there being five energy storage elements in the network, the system is nonetheless of the fourth order, because not all of the elements are independent. Note, for example, that specifying the currents in two of the inductors automatically determines that flowing in the third, by Kirchhoff's current law. Thus, the three inductors actually contribute only two degrees of freedom, diminishing by one the order of the overall network.

FIGURE 20. Equivalent network of coupled *LC* resonators for common-mode initial conditions



The common-mode resonant frequency is thus that of a simple parallel *LC* network:

$$\omega_{cm} = \frac{1}{\sqrt{\left[\frac{(1-k)L}{2} + kL\right]2C}} = \frac{1}{\sqrt{(1+k)LC}}. \tag{56}$$

There are two conjugate imaginary poles of this frequency, so we only need to find the other two poles of this fourth-order network.

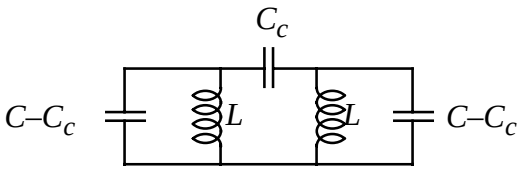
Since a common-mode initial condition is so fruitful in discovering two of the poles, it seems reasonable to try a differential initial condition next. Specifically, if one capacitor voltage is initially made equal to some voltage *V,* and the other to *–V,* (anti)symmetry allows to assert that, however the system state evolves from this initial condition, it must do so in a manner that guarantees zero voltage across the common shunt inductance of value *M*. Consequently, no current flows through it, and the shunt inductance may be removed (either by open- *or* short-circuiting it; both actions will lead to the same answer). Removing that inductance yields the following differential-mode resonant frequency:

$$\omega_{dm} = \frac{1}{\sqrt{[2(1-k)L]\frac{C}{2}}} = \frac{1}{\sqrt{(1-k)LC}}. \tag{57}$$

Now that we've found the pole frequencies, let's see what intuition may be extracted from the exercise. First consider extremely loose coupling, i.e., values of *k* very near zero. In that situation the two mode frequencies are nearly the same, because we have two nearly independent and identical tanks. As *k* increases, however, one resonant frequency decreases, while the other increases; *mode splitting* occurs. The stronger the coupling, the wider the separation in resonant frequencies.

As an illustration that mode-splitting is an extremely general consequence of coupling resonators together, consider the use of capacitive coupling:

FIGURE 21. Capacitively coupled resonators



Here, the individual resonator capacitances are arbitrarily expressed as a function of the coupling capacitance. One could just as well have labeled the resonator capacitances simply as C, but the choice shown simplifies the analytical expressions somewhat, as will be seen.

Following an approach analogous to that used to analyze the magnetically coupled case, we find that the two mode frequencies are given by:

$$\omega_{cm} = \frac{1}{\sqrt{(C - C_c)L}} = \frac{1}{\sqrt{(1-k)LC}} \tag{58}$$

and

$$\omega_{dm} = \frac{1}{\sqrt{(C + C_c)L}} = \frac{1}{\sqrt{(1+k)LC}}. \tag{59}$$

For these equations, an explicit expression for the coupling coefficient, k, is found to be

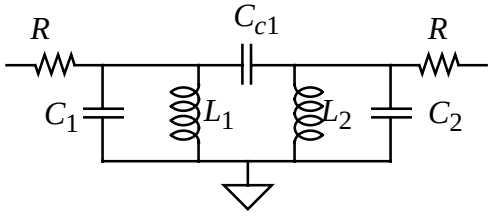
$$k = \frac{C_c}{C}. \tag{60}$$

As with the magnetic case, the coupling coefficient cannot exceed unity (if negative element values are disallowed) when expressed in this manner. As we can see, both magnetic and capacitive coupling give rise to the same splitting of modes. This mechanism is so general that it explains a host of phenomena, such as the formation of energy bands in solids (here, the initially identical mode frequencies – energy levels – of free atoms split as the atoms are brought closer together to form a solid).

From Eqn. 58 and Eqn. 59, it should be clear that one may use a measurement of the two mode frequencies to determine k experimentally. Indeed, for small values of coupling, the difference in mode frequencies (normalized to their geometric mean) is approximately equal to k.

The mode splitting that accompanies coupling permits placement of poles to produce response shapes such as Butterworth and Chebyshev. Simply adding an input and output port to the basic structure of Figure 21, for example, readily produces a bandpass filter which may be extended to any number of stages:

FIGURE 22. Coupled resonator filter



From the analysis of mode splitting, it should seem reasonable that small amounts of coupling (small values of coupling capacitance in this particular example) produce narrowband filters, and that relatively large amounts of coupling produce broadband filters. When this idea is placed on a quantitative basis, it is possible to express the bandpass filter design problem entirely in terms of coupling coefficients, uncoupled resonant frequencies, and tank loading. This reformulation in terms of invariant parameters (e.g., resonant frequency, impedance levels, bandwidth) facilitates the design of microstrip filters where, owing to their distributed nature, it is not always possible to identify individual lumped elements.

Microstrip Filters

1.0 Background

There are two broad classes of distributed filters. One derives from lumped prototypes, and the other doesn't.[1] For the former class, a straightforward synthesis recipe developed for some of the earliest microwave filters still works well for many applications: Simply replace the discrete inductors and capacitors of a lumped prototype with transmission line sections.[2] As discussed in the chapter on microstrip, transmission lines approximate well the behavior of lumped elements if the sections are a suitably small fraction of an electrical wavelength in extent. A short section of open-circuited line functions well as a capacitor, while a short piece of shorted line behaves as an inductor.

One important consideration to keep in mind, however, is that there is always a frequency above which these pieces of line cease to be very short relative to a wavelength. The attendant impedance variation alters the filter response. For example, a microstrip low-pass filter may have a response that pops back up again above the nominal stopband. Since such spurious responses are hardly unique to low-pass filters, one must evaluate carefully any proposed realization to assure that all spurious responses are benign in magnitude or location.

To see how distributed filters may derive from lumped prototypes, recall that the input impedance of a short piece of open-circuited line is approximately

$$Z \approx \frac{Z_0}{j\omega\,(l/v)}, \tag{1}$$

so that its equivalent capacitance is

$$C = \frac{l}{vZ_0} = \frac{l\sqrt{\varepsilon_{r,\,eff}}}{cZ_0}. \tag{2}$$

One can expect about 1.3pF/cm with 50Ω lines on FR4.

Similarly, for the inductance of a short line terminated in a short circuit, we have

$$L = \frac{lZ_0}{v} = \frac{lZ_0\sqrt{\varepsilon_{r,\,eff}}}{c}. \tag{3}$$

1. U. S. President Ulysses S. Grant is said to have quipped, "I know two tunes. One is 'Yankee Doodle,' and the other isn't."

2. See, e.g., Chapter 10 of *Microwave Transmission Circuits*, MIT Radiation Laboratory Series, vol. 9, McGraw-Hill, 1948.

As we've often cited, a typical value for inductance is 1nH/mm for narrow (high impedance) lines.

Keep in mind that these equations also apply approximately even when these line segments are not terminated in perfect open or short circuits. The foregoing equations remain reasonably accurate as long as the segments are terminated in impedances that approximate opens or shorts in comparison with the characteristic impedance of the lines. Hence we would want to choose Z_0 as low as possible (or practical) to make a capacitor, and Z_0 as high as possible to make an inductor.

One cannot specify arbitrarily high characteristic impedances, of course, because there is always a lower bound on the width of lines that may be fabricated reliably. Assuming a typical manufacturing tolerance of 2mils (50μm), and supposing that this variation is allowed to represent at most 20% of the total width, one may assume a minimum practical linewidth of about 10mils (250μm). Hence, on 1/16" FR4, practical line impedances rarely exceed about 200Ω.

There are also practical bounds on the maximum width of the lines because, again, all linear dimensions of a microstrip element must be small compared to a wavelength at all frequencies of interest for close approximation to lumped element behavior. The associated implicit lower bound on impedance depends on the operational frequency range, but as a general rule, characteristic impedances below approximately 10-15Ω are rarely used. In realizing microstrip filters, then, it's important to keep in mind that practical impedance levels are thus generally within about a factor of four of 50Ω.

2.0 Stepped-impedance filters

One extremely simple method for transforming discrete prototypes into microstrip form uses only the narrowest and widest lines that may be comfortably (or repeatably) accommodated. The narrow lines implement series inductors, and the wide lines implement shunt capacitors. Lengths are adjusted to produce the desired component values. As one might expect, the fundamentally approximate nature of the transformation limits its utility. Stepped-impedance filters are thus used in applications where one may tolerate relatively large errors relative to the lumped filter prototype's response. These errors generally increase in significance as one moves higher in frequency. At and below the design cutoff frequency, the stepped impedance and lumped-parameter filters might behave similarly. Beyond cutoff, however, the stepped impedance filter generally fails to roll off as quickly as the prototype and, indeed, the attenuation may never exceed a certain level. Furthermore, the filter's response may exhibit numerous spurious passbands.

That said, let's examine how we may produce a low pass filter using the stepped impedance architecture. As a specific example, assume that we desire a cutoff frequency of 1GHz, and that we use a constant-*k* prototype as the basis for the microstrip filter. If the prototype has two complete T-sections, the stepped impedance filter will have seven segments. Assume further that the minimum and maximum line impedances are 15Ω and

200Ω. To match the lumped element values of the prototype, we require an inductance of 15.915nH, which we implement with the narrowest available line, of a length given by

$$l = \frac{vL}{Z_{0,\,max}} = \frac{f\lambda L}{Z_{0,\,max}}, \tag{4}$$

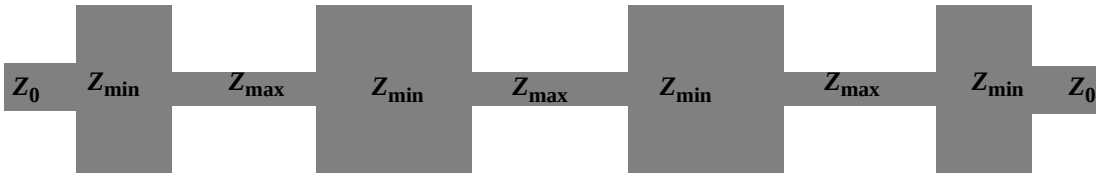
which works out to a normalized length for the inductor of about 28.647° at the cutoff frequency.[3] Similarly the main 6.3662pF capacitors should have a length

$$l = vCZ_{0,\,min} = f\lambda CZ_{0,\,min}, \tag{5}$$

which corresponds to a normalized length of about 28.742°.The half-section terminating capacitors are exactly half that length.

The layout of the filter appears approximately as follows:

FIGURE 1. Stepped impedance filter example (not drawn exactly to scale)



Simulations of this filter show 3.9dB of attenuation at the design cutoff frequency, and a normalized –3dB bandwidth of 0.989. Compare those values with the constant-*k* prototype's 10dB attenuation at the cutoff frequency, and normalized –3dB bandwidth of 0.934. Perhaps more important than those differences is the existence of spurious passbands at around 5.4 and 5.8GHz for this particular implementation. The constant-*k* filter, of course, theoretically exhibits no such spurious passbands, and this difference in behavior must be taken into account in any practical implementation of distributed filters, stepped impedance or otherwise.

3.0 Commensurate-line filters

From Eqn. 2 and Eqn. 3, we see that both the line length and characteristic impedance may be varied to produce a desired inductance or capacitance. The stepped impedance filter arbitrarily uses just two fixed values of line impedance, and varies the length as necessary. An alternative method that is frequently used, but which is ultimately no less arbitrary, involves the use of *commensurate lines*.[4] The term refers to the equality of line lengths used to implement the filter elements. As with the stepped impedance filter, short segments of shorted line implement inductors, and short pieces of open-circuited line act as capaci-

3. Matching the impedances at the cutoff frequency is an arbitrary, but good, choice since behavior in the vicinity of cutoff is often of great importance.

4. P. I. Richard, "Resistor-Transmission Line Circuits," *Proc. IRE*, v. 36, pp. 217-220, Feb 1948.

tors. In Richard's original description of the method (see Footnote 2), "short" is specifically taken to mean an eighth of a wavelength at the cutoff frequency. That is, each inductor or capacitor of a lumped prototype is replaced by a $\lambda/8$ length of transmission line whose characteristic impedance is varied to produce the desired component value. For this particular choice the resulting filter response is periodic in frequency, and may be considered the result of aliasing the lumped prototype's response.

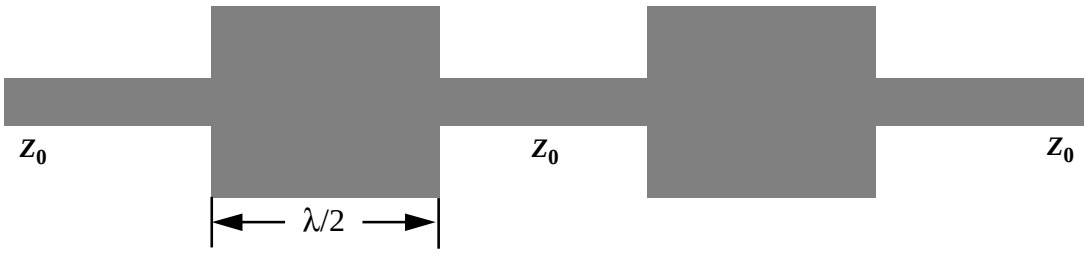
One practical consideration is that following Richard's prescription precisely requires the use of shorted lines to implement inductances. However it is often inconvenient to implement shorts in microstrip, because one would generally like to avoid the use of grounding vias if at all possible. Fortunately, we may again use transmission line behavior to transform inductors into capacitors, and thus avoid the need for shorted sections. Specific examples of how to exploit this idea will be presented in a future note.

4.0 Bandpass filters

4.1 Half-wave filters

The halfwave bandpass filter is actually a low-pass filter, in which what would normally be considered a spurious passband is used as the main filter band. It resembles the stepped impedance filter in some superficial ways, but the element lengths are no longer constrained to small fractions of a wavelength. Indeed, the element lengths are specifically selected to equal a half wavelength at the center of the intended passband:

FIGURE 2. Half wavelength bandpass filter



Recall that a transmission line reproduces at its input the load impedance, whenever the line is an integer multiple of a half-wavelength in extent. For the structure shown, this condition implies a driving point impedance that is equal to Z_0, implying maximum power transfer into the filter and, ultimately, to the load. At frequencies where the electrical length of the filter sections differs significantly from the half wavelength condition, power is coupled into (and out of) the filter less efficiently. From this description, it is clear that the filter behaves essentially as a bandpass filter, with periodically disposed passbands. Note that zero is an integer, so the lowest-order passband is actually centered about zero frequency. Inspection of the physical structure leads to the same conclusion, that the filter indeed passes DC, as asserted at the beginning of this section.

The quality of off-center rejection depends on a mismatch of impedances between the line and the filter section. This observation implies that better stopband rejection should be obtainable if each half-wavelength section is made as wide as possible. Furthermore, the stopband rejection increases with the number of half wavelength sections.

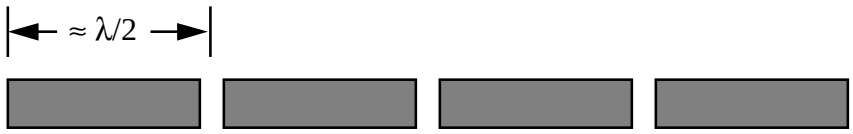
The width of the passband is also a function of the lengths of the line between the filter sections.

4.2 Coupled resonator filters

We've already seen that both electrostatic and magnetic coupling are equally effective for splitting modes. This observation forms the basis for a class of bandpass filters known as coupled-line filters. The underlying idea can be, and has been, implemented in a great many forms, and it is simply impossible to do more than survey a small number of them. The comprehensive work in this field is by Matthaei, Young and Jones (simply referred to by microwave cognoscenti as MYJ), and is a must for anyone who is serious about the subject of microwave filters and impedance matching.[5] Unfortunately, this tome exists only in one edition, and thus does not cover advances made in the last several decades. In particular, microstrip implementations are not extensively covered, so key quantitative design information is often absent.

A good conceptual (but not necessarily practical) starting point is a simple structure based on the lumped prototype bandpass filter, in which the coupling capacitances are implemented by simple gaps in a line:

FIGURE 3. Capacitively coupled microstrip bandpass filter



The segments of line act as half-wave resonators (to avoid the grounded connections that quarter-wave sections would require), and the widths of the interline gaps control the coupling between the resonators. Note, as implied in the figure, the resonator sections are laid out shorter than a half-wavelength, as a consequence of fringing. The larger capacitance of smaller gaps leads to greater coupling, and thus larger bandwidths. For typical FR4, each resonator section is generally on the order of 3" long at 1GHz.

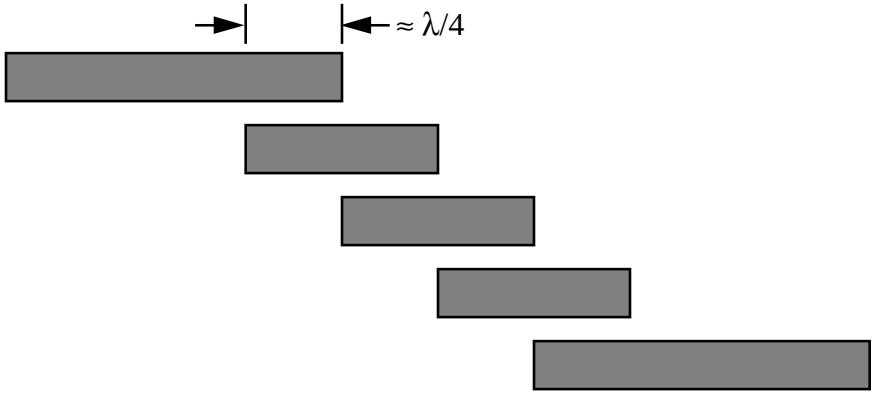
Aside from the difficulty of computing the dimensions required to produce a given level of coupling, there is an additional problem: The required gaps are generally quite small for typical filters. As a result, reproducible filter behavior demands tight control of dimensions.

5. G. L. Matthaei, L. Young and E. M. T. Jones, *Microwave Filters, Impedance-Matching Networks and Coupling Structures*, McGraw-Hill, New York, 1964. It should be mentioned that this work also contains important contributions by Seymour Cohn, who did extensive work on coupled line filters, among others.

Coupling the resonators laterally, rather than simply end-on enables considerable relaxation of dimensional tolerances. Although the increased overlap may imply that the coupling is no longer necessarily purely capacitive, we've already seen that both electrostatic and magnetic coupling mechanisms are equally effective at producing the desired mode splitting. Thus even though the quantitative analysis of coupled lines is somewhat complicated, the intuitive ideas underlying their operation is straightforward.

A simple (as well as simplified), but practical implementation of this idea is shown in the following figure. As seen, the resonators are nominally a half wavelength in extent, and overlap each other by a quarter wavelength.[6] The detailed response shape is controlled by properly choosing the amount of overlap and the characteristic impedances of the lines (for simplicity, all lines are shown of equal impedance level in the figure). To achieve the same effective coupling as in the capacitively coupled case, the greater overlap must be compensated by a larger line-to-line spacing, easing the tolerances as advertised.

FIGURE 4. Coupled line microstrip bandpass filter



This type of bandpass filter works quite well, and is quite widely used as a result. One criticism, however, is that the filter occupies a relatively large area (or at least it occupies an irregularly shaped region), especially if the filter uses many sections. In many cases, the layout of the resonator array is rotated, and the input/output couplings suitably bent, to fill better a rectangular space.

A popular alternative modification instead folds each individual resonator into a hairpin shape, producing a more compact design:

6. The physical line lengths must be adjusted downward in practice to accommodate end fringing, just as in the filter of Figure 3.

FIGURE 5. Hairpin bandpass filter

Note that, as in other bent transmission lines, the resonator sections all use mitered bends to minimize unwanted discontinuities.

Aside from the difficulty of tuning individual stages properly, spurious modes are the bane of filter designers. Many practical filters may exhibit poor attenuation to certain signals that are nominally in the stopband. These spurious responses may be a fundamental property of the filter architecture (e.g., the natural periodicity of the impedance of a piece of terminated line, or the excitation of surface or higher-order modes in a microstrip section), or may result from parasitic elements associated with components used to build the filter.

5.0 Summary

We've seen that modern filters trace their lineage all the way back to Heaviside's hundred-year old analysis of telegraph transmission lines, with important contributions by Campbell, Wagner and Zobel shortly thereafter. The inability to specify detailed response shapes of such wave filters motivated the development of the Butterworth, Chebyshev and Cauer approximations.

Sensibly enough, coupled low-pass sections are used to make low-pass filters, and coupled resonators can be used to make bandpass filters. These observations hold even when the individual sections are realized with distributed (or quasi-distributed) elements, leading to a host of filter implementations that are amenable to realization in planar form. Examples of these include the parallel-coupled, interdigitated, and hairpin microstrip filters.

Antenna and Filter Design Lab

1.0 Introduction

In this lab experiment, you will design, construct and test two types of antennas intended for use at 1GHz. The two kinds you will build are an ordinary monopole and a patch antenna. It is particularly important to play around with them and see how sensitive the impedance characteristics are to proximity to objects. Also pay attention to the accuracy of the formulas presented. Do the lab's physical characteristics satisfy the assumptions that implicitly underlie the equations?

You will also build and test a simple bandpass filter.

Reading assignment: Handouts #13-15 (Antennas, Filters and Microstrip Filters); due date: Friday, May 4. This is a lot of reading if you try to understand every word, so don't do that. Instead, do a quick skim to see what's there, then focus on the parts that are relevant for this lab.

2.0 Monopole

The key design equation for a monopole over a ground plane is:

$$R_r \approx 40\pi^2 \left(\frac{l}{\lambda}\right)^2 . \tag{1}$$

In this experiment, the ground plane will be imperfect (as is generally the case). Specifically, you will continue to use the pre-cut pieces of FR4 as the ground plane (yes, very cheesy; in general, one would like a ground plane that extends at least a quarter wavelength away from the antenna in all directions). The antenna itself is just a stiff piece of wire jammed into a BNC mounted on your board. If we can obtain some unassembled male BNCs in time, you may be able to build the antenna with one of those. However, don't count on this happening.

Do NOT jam this wire into any connector attached to the network analyzer. Also, because it will be altogether too easy to poke your (or someone else's) eye with such an antenna, put a ball of black electrical tape (or some other such contrivance) on the end. *****Email Sergei that you have read and understood both of these instructions.***** We want to make sure that everyone finishes EE414 with as many functioning eyeballs as at the beginning of the term, and that the network analyzer suffers a minimum of additional trauma.

You will face a small mechanical engineering challenge (this is part of the experiment), deciding where to mount the connectors, as well as choosing the length of the microstrip feedline that goes from one connector to the other. Note that zero is a possible length.

Cut the monopole to produce a good match at 1GHz. Plot out S11 over both a broad frequency range (e.g., octaves), and over a narrow one (below an octave, centered about 1GHz). Report the "impedance bandwidth," which we will define here as the frequency range over which the SWR is under 2. This is a common, arbitrary and very generous specification. Because such a large SWR is intolerable in many applications, also report the range over which the SWR is below 1.5 and 1.2.

You will note a sensitivity to the proximity of objects, so you will have to decide on standard test conditions (e.g., monopole pointing up, partner facing east, you lying supine, precisely π meters away).

After you've made these measurements, experiment with moving your hand near the antenna, observing how the impedance changes. Feel free to bring other objects near the antenna. Try conductors and dielectrics. Can you say anything ~quantitative about how close an object must get before it has a "significant" effect on the impedance? How does this measured distance compare with a wavelength? Can you explain this distance in terms of physics? Can you say anything about the nature of the perturbation (i.e., whether it increases capacitance or inductance)?

3.0 Microstrip Patch Antenna

A half-wave patch antenna at 1GHz is a little too long to implement using our pre-cut FR4 boards. However, a quarter-wave antenna is not out of the question. Using the formulas for effective electrical length, compute the physical length of the antenna (it should work out roughly to about 1.4-1.5 inches). If you lay this out along one of the long edges of the board, it will be easy to produce an excellent ground by merely folding a length of copper foil tape over the edge and soldering.

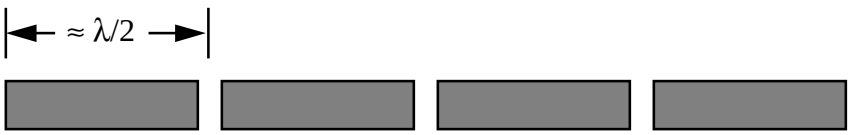
As to width, you will have to experiment to find the value that result in a good match. Given the development in the notes, one might expect the width to be between 2 and 3 times the length. However, your mileage may vary.

After achieving a good match, fool around with it in the same way as the monopole. Characterize the impedance bandwidth and sensitivity to nearby objects.

4.0 Capacitively Coupled Bandpass Filter

Design a simple bandpass filter centered at 1GHz, of the following type:

FIGURE 1. Capacitively coupled microstrip bandpass filter



To simplify things, use only one half-wave section of 50Ω impedance. Initially cut it to your best estimate of exactly half a wavelength, knowing that you'll have to cut some pieces off to bring the center frequency to 1GHz.

The –3dB bandwidth target here is about 200MHz. It should be no less than 100MHz, and no greater than 300MHz. This is a loose enough range that you should not have to spend an inordinate amount of time to achieve satisfactory performance. The main trickiness here is simply getting the coupling gaps to the right value to produce the desired response.

Once you've designed the filter, note the insertion loss (i.e., the loss at the center of the passband), and compare to the 0.08dB/inch attenuation figure of a pure line. Also, note the location and severity of several spurious passbands. In particular, note that the filter acts like a bandpass filter at lower frequencies, but like a high pass filter at high frequencies. You should be able to explain, with a minimum of mathematics, why you see what you see.

Narrowband PA Design Lab

Due date: May 11, 2000

1.0 Introduction

In this lab experiment, you will design, construct and test a single-transistor power amplifier for use at 1GHz. The design goals are: Output power at 1GHz into 50Ω > 50mW (17dBm), power gain >8dB, input and output return losses greater than 10dB, and the highest collector efficiency you can achieve while satisfying these conditions. Even though the 2SC3302 is far from a high power device, it is the transistor you are to use. Because the power is limited, any RF-related biohazards are similarly limited. Despite what you might think, we really do care about your health and safety!

Also, **UNDER NO CIRCUMSTANCES ARE YOU TO GENERATE MORE THAN 200mW OF OUTPUT POWER.** You can actually get these wimpy transistors to produce more than this on a short-term basis, but I don't want you even to try. Furthermore, the network analyzers have an input power limit beyond which they distort, and another beyond which they are damaged. Be sure to observe these limits.

Reading assignment: Chapter 13 (on power amplifiers) in the textbook. If you also feel the need for more knowledge about bipolar transistors, another handout will be made available on request.

2.0 Issues

You have several design degrees of freedom. First, you are free to choose any topology. There is no linearity specification here, so you have more freedom than in most real life PA designs. Another important set of decisions concerns the impedance level seen by the collector. Although 50Ω is the ultimate load impedance, there may be an advantage to providing an impedance transformation. The supply voltage is a variable you should exploit. The decision will be constrained by a knowledge of the maximum voltage and current that the 2SC3302 can tolerate reliably.

Often, producing an input match is a challenge with bipolar power amplifiers. The reason is simple: Power amplifiers are large-signal beasts, and the input diode of a bipolar transistor is a highly nonlinear element under large-signal conditions. In many practical bipolar amplifiers, this problem is handled in a shockingly brute-force fashion: Shunt the base-emitter diode with a suitably low value resistor. With a sufficiently small resistor, the impedance of the parallel combination will be dominated by the linear resistor, at the cost of gain. The small-signal trick of inductive emitter degeneration is inadequate for power amplifiers of any reasonable power level. For the modest specifications you are to meet in this lab experiment, you might be able to get away with some combination of small-signal tricks and resistive shunting.

Synthesizer Design Lab

1.0 Introduction

In this lab experiment, you will design, construct and test a frequency synthesizer-based local oscillator. This is by far the longest and most complicated lab exercise of the entire course, and you have ~2 weeks in which to do everything. Do not wait until a week has gone by to begin the lab, for the amount of material to read and understand is large, as is the amount of labor, even though we are going to provide you with a fair amount of pre-fabricated items so that you don't have to derive everything from first principles. But there is still a lot of reading.

There will be only one more lab exercise in the course, in which you will integrate all of the blocks into a transceiver, so we're getting to the really exciting stuff! A mixer and filter will be added at that final phase, but the design of these will involve minimal additional labor.

The synthesizer consists of two main modules: A VCO and the rest of the PLL. It is advisable for each partner to take ownership of one block's design, so that you may work in parallel. Close collaboration with other groups is also **strongly** encouraged.

The specifications are these: VCO frequency range of <900MHz to >1GHz for the receive LO synthesizer (in the final lab, you will replicate your design for the transmitter, with a slightly different tuning range of <950MHz to >1.05GHz; if you can get enough tuning range, then you will only have to copy a single design); output power of 0dBm at absolute minimum (and preferably 5-7dBm max) for driving a diode ring mixer. All spurs with the loop closed should be at least 30dB below the carrier.

In your writeup, show your linearized PLL loop model, describe how you chose the loop bandwidth, and give component values for your loop filter. Be sure to include a tuning curve for your VCO (i.e., frequency vs. control voltage), and report the nominal, minimum, and maximum values of VCO gain constant over the specified frequency range. Design the PLL to possess a phase margin of at least 45° over the tuning range. In your loop calculations, make sure that you don't confuse radians per second with hertz!

Measure the phase noise at 100kHz and 1MHz offsets from the carrier. Also report the size of the f_{ref} spur, relative to the carrier. Is this the biggest spur? If not, what is, and how big is it? Can you relate the frequency of this spur to "other periodic" signals in the loop?

Reading assignment: Check the relevant sections of the textbook chapters on oscillators, PLLs and synthesizers (Chapters 15 and 16). Those of you with a weak background on feedback may wish to skim the chapter on feedback systems (Chapter 14), or else much of the material on PLL loop filter design will be completely incomprehensible. Please also read the handout on phase noise measurement.

Due date: Friday, May 18, 2001.

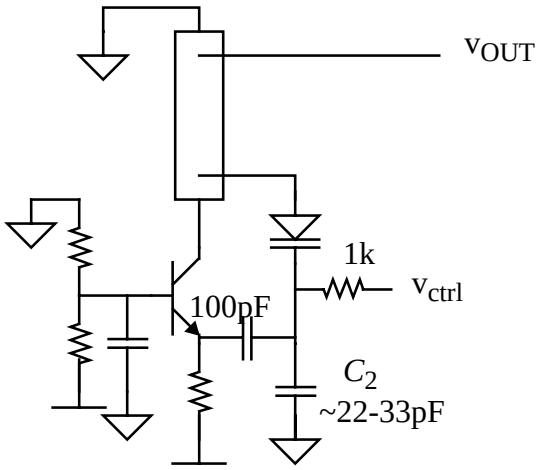
2.0 VCO

I *strongly* recommend a Colpitts topology for the VCO, but you are certainly free to use any topology you wish. If you do use something else, I would be very interested in seeing the results of any such efforts. This handout will focus on a Colpitts implementation, but a "negative resistance" oscillator will also be described briefly in class.

The Colpitts oscillator is just a positive feedback system with a capacitively tapped resonator. Many topological variations fit within that verbal description, but not all are equally amenable to convenient biasing or provide good tuning range. For example, you could ground the collector, or tie it to some positive V_{cc}. Or you could ground the base. Each of these choices leads to a different collection of bias headaches and sensitivities to parasitic-induced problems. You are therefore encouraged to investigate these alternatives some day, after the course has ended and you have a lot of free time.

The following Colpitts implementation has worked well for me (and, perhaps more important, for previous EE414 students). It is offered free, but without warranty (you get what you pay for!):

FIGURE 1. Slightly simplified schematic of VCO (read text!)



First note one significant feature of the circuit: the collector's DC potential is ground; *negative* voltages are used to bias the transistor. **Be sure that you understand this and hook things up right, or else you will blow up parts, and have to re-do enormous amounts of work!** This choice of polarity is driven by several considerations: Returning the tank to ground is nice because it avoids having to bypass a collector supply (with either chokes or BFCs), and also eliminates an output coupling capacitor. All of these simplifications make it easy to terminate one end of the inductor in an excellent short (locate the inductor close to the edge of the board to take full advantage of this topology). Furthermore, the control voltage from the PLL is ground-referenced, so annoying biasing gymnastics are avoided by connecting the varactor to the ground-referenced load structure.

In this configuration, a suitable length of line acts as an inductor (kinda sorta; more on this later). That inductance resonates with a capacitance formed by the series combination of

the varactor, and the sum of C_2 and the emitter capacitance. The tuning range is therefore a function of the varactor capacitance versus that other total capacitance. Fortunately, with the particular arrangement shown, it is possible to design things so that the total capacitance is dominated by that of the varactor. This condition is readily produced by choosing a *large* enough C_2, and is what we want in order to maximize tuning range. Unlike some other Colpitts arrangements, then, parasitic capacitances across C_2 are relatively benign. We can't make C_2 arbitrarily large, however, because it is also part of the feedback voltage divider. If it is too large, the feedback may diminish to the point where oscillation ceases. Here, we have arbitrarily chosen C_2 equal to the maximum capacitance of the varactor. What is this maximum? The PLL chip produces a control voltage that nominally swings from 0.5V up to 4.5V. The MV105 varactors nominally tune 8 to 18pF over that range. I'd have preferred a 4pF to 9pF range, but these are the varactors we have in stock. Many past EE414 students achieved excellent results by using the 2SC3302 emitter-base junction capacitor in place of the MV105, so you may wish to consider this option in your design. If you do, it is advantageous to short the collector to the base (doing so removes an antenna for noise injection, and reduces series resistance somewhat).

Another important consideration is to keep the base bypass path as short as possible. Any inductance in series with the base can result in oddball parasitic oscillations at frequencies you never imagined could exist!

2.1 Load inductor design

Assuming a net capacitance of 8pF at the center of the tuning range (calculated by using the geometric mean of what you get when you look at 8pF, then 18pF, in series with 22pF+4pF of emitter cap and strays[1]), you'd want an inductance of about 3nH. This is going to be a little tricky, but not impossible, to implement. You'll need to lay things out rather carefully. In particular, the stray inductance of the varactor connection must be absolutely minimized! If you succeed, the calculated tuning range will be about ±15% about the nominal frequency, comfortably in excess of the ±5% that you are asked to provide. Inevitably some other factors conspire to reduce the range below the calculated value, but you should still end up with enough margin here.

To ease the inductor design problem and simultaneously increase tuning range, you can increase the voltage swing across the varactor to reduce its minimum capacitance. However, this means that you will have to design and build an amplifier to take the control voltage from the PLL chip, and gain it up appropriately. So, you trade complexity in one domain for complexity in another. Pick your poison.

2.2 Bias

From the chapter on oscillators, you know that the oscillation amplitude is a function of bias current and losses (which include the action of the load resistance). The last item is

1. This is a pessimistic estimate of total "other" capacitance; your mileage may vary, especially if you use a transistor in place of the varactor.

under your control to a certain extent through choice of output tap location. The shortness of the load inductor limits the ease with which you can select and vary the tap location, so bias current will be an important design variable for you. Depending on where you end up tapping the output, a 0 to 7dBm output will probably result with bias currents in the range of low to medium numbers of milliamps. Be sure that, in laying out the bias network, you avoid stray capacitance where it would affect the frequency of the tank, or tuning range could degrade severely. As with the tapped LNA tank, this load structure doesn't guarantee control over both the real and imaginary parts of the output impedance, so additional matching may be required.

Another consideration is that the anode of the varactor may see a zero DC voltage, but there will also be an AC component, too. The control voltage needs to be sufficiently larger than the peak anode voltage to avoid forward-biasing the varactor. Oscillation might even cease if this problem occurs, causing the loop to get confused. There is therefore a constraint on the tank amplitude. So, don't just apply a zillion amps to the oscillator thinking that bigger is automatically better.

The simplified schematic shows no values for the biasing components; these are left for you to design. Bias stability is not specified in this lab, but choose values that lead to "reasonable" stability, just as for the LNA design. As a starting point, try dropping a volt or two across the emitter resistor, and selecting the base bias dividers to carry a current of approximately the collector current divided by 10. Fortunately, biasing this oscillator involves less work than biasing the LNA, because we don't have to arrange for feedback from the collector. Here, we're allowed to put junk into the emitter with relative impunity.

2.3 "It doesn't work; now what do I do (after weeping uncontrollably)?"

The tank inductor can be the source of difficulties because of its shortness. You might have reasoned that, the narrower the line, the better the tuning range, because a shorter line could then be used for a given impedance, leading the line to act more like a pure inductor than a quasi-resonator. After all, resonator loss is only a second-order function of width because most of the loss in FR4 is due to the dielectric, as mentioned in the notes on microstrip, so using narrow lines may not be so unattractive here.

However, shortness is a problem for largely mechanical reasons. To allow a somewhat longer line to be used, you might consider a wider line than you might normally try. Just make sure that all dimensions remain well below a quarter wavelength, or else strong distributed effects will throw off your calculations. It is preferable in real designs to choose a length that will not produce high impedances at some multiple(s) of the oscillation frequency, because those distortion products will not be filtered out by the tank (remember, the "inductor" you've made is in reality a transmission line that is terminated in a short). In this lab experiment, however, there is no distortion specification, so you don't have to obsess about this issue here. It's just mentioned here so that when you go out into "the real world," you can't rightly complain that some ivory tower academic never told you about this issue.

Another mildly tricky issue is that the tuning range is narrow enough, and the varactor tolerances loose enough, so that the center frequency of the VCO may not be quite the value you want. If you're close to the right frequency, you may not have to rip out the inductor and start over. To raise the frequency a little bit, add some copper near the ground end of the line to decrease inductance. To decrease frequency a tiny bit, add a solder blob (the closer to the collector, the greater the effect), or narrow up the line by slicing a little piece out of it. To decrease frequency further, solder a short length of wider foil onto the line to increase capacitance (again, the closer to the collector, the larger the effect).

If your output power level is too low, the main cure is more bias current. If your power varies wildly over the tuning range, the cure is to tap the output from a point closer to the collector end of the line. Why? The loss of the line, and hence its effective resistance, varies with frequency. If this frequency dependent loss dominates, then the amplitude will vary significantly with frequency. Tapping the output closer to the collector loads down the tank more severely, but at least the load is more constant. An increase in bias current can compensate for the drop in average output power. The tradeoff is one of loop gain vs. output power flatness vs. filtering quality (heavier loading implies degraded *Q*).

Careful thought given to layout issues *before* beginning construction will greatly increase your chances of a successful design.

Stay tuned to the class website for additional hints and tips as they develop.

3.0 PLL Synthesizer

3.1 Overview

The overall transceiver uses frequency modulation and a 100MHz IF. You have the option of modulating either the reference oscillator or the VCO inside the transmit PLL itself. If you choose the former, remember that the frequency deviation corresponding to your modulation will be multiplied upward by the same factor as the base reference frequency. Also, you will need to select the PLL bandwidth to exceed the highest modulation frequency you wish to employ, or the synthesizer will filter out the modulation.

On the other hand, if you employ direct FM, in which you modulate the transmit VCO's control voltage directly, you will need to choose the PLL bandwidth well *below* the lowest modulation frequency of interest. You should understand, and be able to explain, why the requirements on loop bandwidth are completely opposite for these two choices. Also note that no multiplication accompanies the use of direct FM. You should be able to explain the consequences of multiplication (or lack thereof) on the received or demodulated signal.

Those preceding two paragraphs apply to the transmitter but, here, you are only asked to design the receiver synthesizer. Those issues are raised here simply to orient your thinking a little bit in advance of final integration.

We are very fortunate that Motorola makes a chip that contains most of the circuitry for building a synthesizer, otherwise you would never finish the lab (you'd spend all of your time designing and wiring up a hulking mass of flip-flops, op-amps and still end up with smoke-emitting semiconductors). The MC12181 contains crystal oscillator circuitry (which will be used to generate a low frequency reference that the chip subsequently multiplies upward by a user-defined value), a four-bit digitally-controlled frequency divider, a phase detector, and a charge pump. The user just has to provide power, a crystal, a VCO and a loop filter to complete the synthesizer. This portion of the lab, then, mainly concerns designing the loop filter, with a little crystal oscillator design thrown in.

The data sheet and applications notes for the MC12181 are attached to this handout. We will provide you with a pre-made PC board that allows you to solder the 12181 and its auxiliary components into "the right places" to save you enormous amounts of time in layout and fabrication. **Be sure** to thank your TA, Moon Kim, for procuring parts, prototyping the synthesizer design, and laying out the PC board artwork! Without her efforts (all performed under quite extreme time pressure), there is no chance that you would have any hope at all of completing the design in finite time.

The schematic of the synthesizer you will build is almost exactly as shown in Figure 2 on page 3 of the data sheet. Exceptions: Vp and Vcc will be tied together, so you can eliminate one pair of bypass capacitors (either the one on pin 3 or 4); and C_1 is replaced by a slightly more complicated *LC* circuit.

As you can see from the data sheet, the four-bit programming nibble allows you to control the multiplication factor from 25 to 40 (control bit D on pin 13 is the MSB, and the signals are active high). The control pins have weak internal pullup resistors (~ 50-100 kilohms), so you don't have to add any external pullups. Grounding the pins to produce a logic zero is perfectly acceptable, as these pins were meant to be driven by standard CMOS gates.

3.2 Reference crystal oscillator

The reference frequency that is multiplied upward by the programmable factors is controlled by the crystal connected across pins 1 and 2, and is allowed to be as high as 25MHz. The minimum guaranteed upper frequency of operation is therefore 1GHz, which is just barely enough for our purposes. Naturally, we have procured a quantity of 25MHz crystals. The oscillator topology inside the 12181 is a traditional Pierce circuit, in which the crystal is placed across the gate and drain of a common-source FET amplifier. The resistor between pins 1 and 2 biases the FET into the active region, and the two capacitors provide additional phase shift beyond what the crystal provides, to satisfy the conditions for oscillation. As mentioned in the textbook, the crystal in a Pierce thus operates at a frequency somewhat above its series resonance, so that it presents a net inductive impedance under normal operation. The exact oscillation frequency depends on both the crystal *and* the two capacitors, although it is much more sensitive to the crystal's resonant frequency (because of the vastly steeper reactance-vs.-frequency curve of crystals, relative to that of capacitors). Each crystal intended for use in a Pierce is specially cut to oscillate on frequency only with a specified capacitive load. We will be using 15pF capacitors.

As mentioned in the textbook, high frequency crystals are very thin, making them fragile and difficult to manufacture. The designers of the 12181 know this, and consequently specify a 25MHz upper frequency, which corresponds to about the maximum frequency at which crystal manufacturers will provide an inexpensive fundamental-mode crystal. Unfortunately, 25MHz is high enough that many manufacturers (including many of our suppliers) prefer to make an 8.33MHz crystal to save cost, and expect the user to operate it on the third overtone. An ordinary Pierce circuit (such as the one in the 12181), unfortunately, may satisfy conditions for oscillation at both the fundamental and overtone frequencies, so a modification must be made to poison the loop conditions at the fundamental mode frequency or else some very weird and undesirable effects may result. The easiest way to accomplish this feat is to add an inductance in series with one of the capacitors to produce a series resonance at the fundamental. This resonance produces a short to ground that reduces loop gain to zero, preventing oscillation at that frequency. To restore proper operation at the third overtone, a capacitor in shunt with the added inductor would have to be provided. For our circuit, the necessary inductance is around 24μH, which is a somewhat large value. We may be able to obtain some physically small inductors of this value in time for the lab, but if not (and this is a distinct possibility), you can always wind your own (it's fun, and good for you besides). There is plenty of enamel-covered "magnet" wire in the lab, and you can use Wheeler's formula for inductance. Just be sure to scrape off the translucent insulation on the ends of the wire (many students mistake the copper-colored insulation for copper itself, and are disappointed at the infinite resistance). Sandpaper is ideal for this purpose, and is provided. If the ends accept solder, then you have successfully removed the insulation.

Moon's inductor design uses 30 turns on a 25mm diameter form (an empty sample-size plastic shampoo bottle in this case). The length of the inductor is about 20mm. After winding the coil, the form may be left in place if it is not conductive or otherwise too lossy. The winding capacitance is large enough that no additional capacitance across it is necessary to make the reference oscillator operate on the third overtone. However, the frequency might not be 1GHz to as many significant digits as is stamped on the crystal. We will not worry about a ~100-200 ppm error in frequency in this lab, but you would have to worry about it in a real product. [The capacitance of a small surface mount inductor may be too small, so you might end up having to add some!]

The feedback biasing resistor across the crystal should be 50 kilohms, according to the data sheet (47k and 56k are the nearest standard 10% values). Its value isn't critical, but higher values yield higher *Q*, and lower ones yield better bias stability.

3.3 Loop filter design

As discussed in the textbook, the loop filter's purpose is to remove the "teeth" produced by the phase detection process (which, if you recall, is fundamentally a sampled system in digital implementations, such as this one). Since the control voltage directly modulates the frequency of the VCO, any AC component of control voltage results in a frequency modulation of the oscillator. If these components are periodic, they produce stationary sidebands (spurs). One obsession of synthesizer designers is the systematic eradication of spurs. Spurs unfortunately arise very easily from noise injected into the control line from

supply noise or from external fields (computer monitors are a notorious source of spur-inducing noise). Your VCOs may possess tuning sensitivities of tens of MHz per volt, so even a few millivolts of noise will generate noticeable spectral artifacts.

For a given loop bandwidth, a higher order filter provides more attenuation of out-of-band components. However, the higher the order, the more poles there are, and the more poles there are, the harder it is to make the loop stable. For this reason, many cheesy synthesizer loops are second order, but these rarely provide competitive performance.

Note that the data sheet circuit suggests a loop filter that contains three capacitors. Remembering that the VCO adds another pole (at the origin), we see that the resulting loop dynamics are fourth order (two too many for cheesy engineers!). In the past, designing such a filter involved staring at lots of plots since no simple closed-form design method existed. The result was that most designs were suboptimal because the labor required to find an optimum was simply too great. Luckily this situation has changed quite recently, thanks to the Ph.D. work of Stanford student Hamid Rategh. The following cookbook procedure for a near-optimal loop filter design are those of Hamid.

Step 1: Specify a phase margin. Once this value is chosen, it sets a constraint on capacitor values. Specifically,

$$PM \approx \text{atan}\left(\sqrt{b+1}\right) - \text{atan}\left(\frac{1}{\sqrt{b+1}}\right), \tag{1}$$

where “atan” is Framespeak for “arctangent,” and

$$b = \frac{C_0}{C_A + C_X}. \tag{2}$$

Choosing a phase margin of 50° to provide a little breathing room above the specified value of 45° minimum, we find that *b* should be about 6.5.

Step 2: Select loop crossover frequency. Combined with the results of Step 1, we find the location of the loop stabilizing zero as follows:

From the textbook discussion, we know that maximizing loop bandwidth maximizes the frequency range over which the presumably superior phase noise characteristics of the crystal oscillator are conferred on the output. Unfortunately, the loop is a sampled data system, and we can only push up the crossover frequency to about a tenth of the phase comparison frequency before we start to degrade phase margin seriously. In the 12181, the phase comparison frequency is one-eighth the reference frequency because the reference oscillator’s output is first divided by a fixed factor of eight before feeding the phase detector. Assuming operation at 1GHz with a reference frequency of 25MHz, we find that the phase comparison frequency is therefore 3.125MHz. Choosing a crossover frequency of 100kHz is well below the danger point, so let’s use that value in what follows (you are free to choose some other value, within limits). Hamid says that

$$\omega_c \approx \frac{\sqrt{b+1}}{\tau_z} = \frac{\sqrt{b+1}}{R_0 C_0}. \tag{3}$$

For our numbers, τ_z works out to 4.38μs.

Step 3: Calculate C_0, the value of the zero-making capacitor.

$$C_0 = \frac{I_P}{2\pi}\frac{K_0}{8N}\sqrt{\frac{b+2}{1+\frac{1}{b+1}}}\frac{b}{b+1}\frac{1}{\omega_c^2}, \tag{4}$$

where I_P is the charge pump current (nominally 2.2mA for this part) and K_0 is the VCO gain constant in radians per second per volt.

Step 4: Calculate $R_0 = \tau_z/C_0$. This completes the design of the main part of the loop filter.

Step 5: Select $\tau_x = R_X C_X$ within the following range:

$$0.01 < \frac{\tau_x}{\tau_z} < 0.1. \tag{5}$$

Within these wide limits is considerable freedom of choice. You can choose to design for the arithmetic mean, or the geometric mean, or some other kind of mean. Typically, one selects τ_x to be 1/30 to 1/20of τ_z. A bigger time constant results in somewhat better filtering action, but tends to be associated with lower stability. Since loop constants aren't constant, it is prudent to design for some margin.

Step 6: Complete the remaining calculations.

Back in Step 1, we developed a constraint on the capacitance ratios. Having found one of the capacitances, we now know the sum of C_A and C_X. You are free to select the individual values over a quite wide range, as long as they sum to the correct value. Arbitrarily setting them equal is a common choice.[2] Having done so then allows us to determine their absolute values, which subsequently allows us to determine the value of R_X.

This completes the design of the loop filter.

As with the VCO, be sure to check the website and email frequently for updates and hints.

2. The noise generated by the resistors in the filter will produce broadband modulation of the VCO, resulting in phase noise. Minimizing the phase noise would impose additional constraints on the loop filter design, but complicates the situation enough that the cookbook procedure offered here is all we'll consider.

Order this document by MC12181/D

MOTOROLA

MC12181

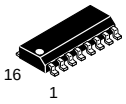
125-1000 MHz Frequency Synthesizer

125 – 1000 MHZ FREQUENCY SYNTHESIZER

SEMICONDUCTOR TECHNICAL DATA

The MC12181 is a monolithic bipolar synthesizer integrating a high performance prescaler, programmable divider, phase/frequency detector, charge pump, and reference oscillator/buffer functions. The device is capable of synthesizing a signal which is 25 to 40 times the input reference signal. The device has a 4–bit parallel interface to set the proper total multiplication which can range from 25 to 40. When combined with an external passive loop filter and VCO, the MC12181 serves as a complete PLL subsystem.

- 2.7 to 5.5 V Operation
- Low power supply current of 4.25 mA typical
- On chip reference oscillator/buffer supporting wide frequency operating range from 5 to 25 MHz
- 4–bit parallel interface for programming divider (N = 25 .... 40)
- Wide 125 – 1000 MHz frequency of operation
- Digital phase/frequency detector with linear transfer function
- Balanced Charge Pump Output
- Space efficient 16 lead SOIC package
- Operating Temperature Range of –40 to 85°C
- > 1000 V ESD Protection (I/O to Ground, I/O to V_{CC})



D SUFFIX
PLASTIC PACKAGE
CASE 751B
(SO–16)

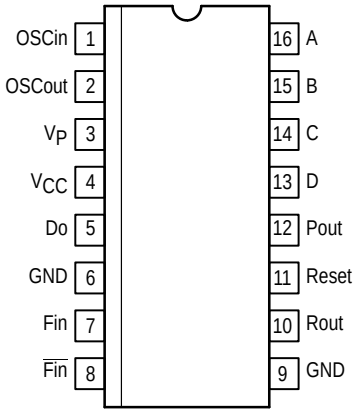
The device is suitable for applications where a fixed local oscillator (LO) needs to be synthesized or where a limited number of LO frequencies need to be generated. The device also has auxiliary open emitter outputs (Pout and Rout) for observing the inputs to the phase detector for verification purposes. In normal use the pins should be left open. The Reset input is normally LOW. When this input is placed in the HIGH state the reference prescaler is reset and the charge pump output (Do) is placed in the OFF state.

The 4–bit programming interface maps into divider states ranging from 25 to 40. A is the LSB and D is the MSB. The data inputs (A,B,C, and D) are CMOS compatible and have pull–up resistors. The inputs can be tied directly to Vcc or Ground for programming or can be interfaced to an external data latch/register. Table 1 below has a mapping of the programming states.

Table 1. Programming States

D	C	B	A	Divider
L	L	L	L	25
L	L	L	H	26
L	L	H	L	27
L	L	H	H	28
L	H	L	L	29
L	H	L	H	30
L	H	H	L	31
L	H	H	H	32
H	L	L	L	33
H	L	L	H	34
H	L	H	L	35
H	L	H	H	36
H	H	L	L	37
H	H	L	H	38
H	H	H	L	39
H	H	H	H	40

PIN CONNECTIONS

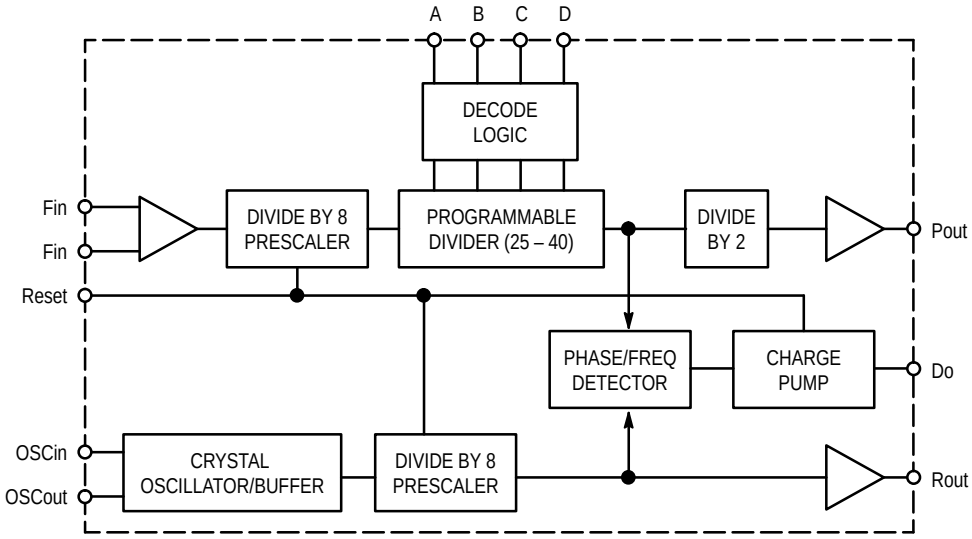


(Top View)

ORDERING INFORMATION

Device	Operating Temperature Range	Package
MC12181D	T_A = –40° to +85°C	SO–16

Figure 1. MC12181 Programmable Synthesizer



PIN NAMES

Pin No.	Pin	Function
1	OSCin	An external parallel resonant, fundamental crystal is connected between OSCin and OSCout to form an internal reference crystal oscillator. External capacitors C1 and C2 are required to set the proper crystal load capacitance and oscillator frequency (Figure 2). For an external reference oscillator, a signal is ac–coupled into the OSCin pin. In either mode a 50 kΩ resistor MUST be connected between OSCin and OSCout.
2	OSCout	Oscillator output, for use with an external crystal as shown in Figure 2.
3	V_P	Positive power supply for charge pump. V_P MUST be greater than or equal to V_{CC}. Bypassing should be placed as close as possible to this pin and be connected directly to the ground plane.
4	V_{CC}	Positive power supply. Bypassing should be placed as close as possible to this pin and be connected directly to the ground plane.
5	Do	Single ended phase/frequency detector output. Three–state current sink/source output for use as a loop error signal when combined with an external low pass filter. The phase/frequency detector is characterized by a linear transfer function.
6	GND	Ground. This pin should be directly tied to the ground plane.
7	Fin	Prescaler input – The VCO signal is ac–coupled into the Fin Pin.
8	Fin	Complementary prescaler input – This pin should be capacitively coupled to ground.
9	GND	Ground. This pin should be directly tied to the ground plane.
10	Rout	Open emitter test point used to verify proper operation of the reference divider chain. In normal operation this pin should be left OPEN.
11	Reset	Test pin used to clear the prescalers (Reset = H). When the Reset is in the HIGH state, the charge pump output is disabled. The Reset input has an internal pulldown. In normal operation it can be left open or tied to ground.
12	Pout	Open emitter test point used to verify proper operation of the programmable divider chain. The output is a divide–by–2 version of the programmable input to the phase/frequency detector. In normal operation this pin should be left OPEN.
13 14 15 16	D C B A	Digital control inputs for setting the value of the programmable divider. A is the LSB and D is the MSB. In normal operation these pins can be tied to V_{CC} and/or ground to program a fixed divide or they can be driven by a CMOS logic level when used in a programmable mode. There is an internal pull–up resistor to V_{CC} on each input.

Figure 2. Typical Applications Example

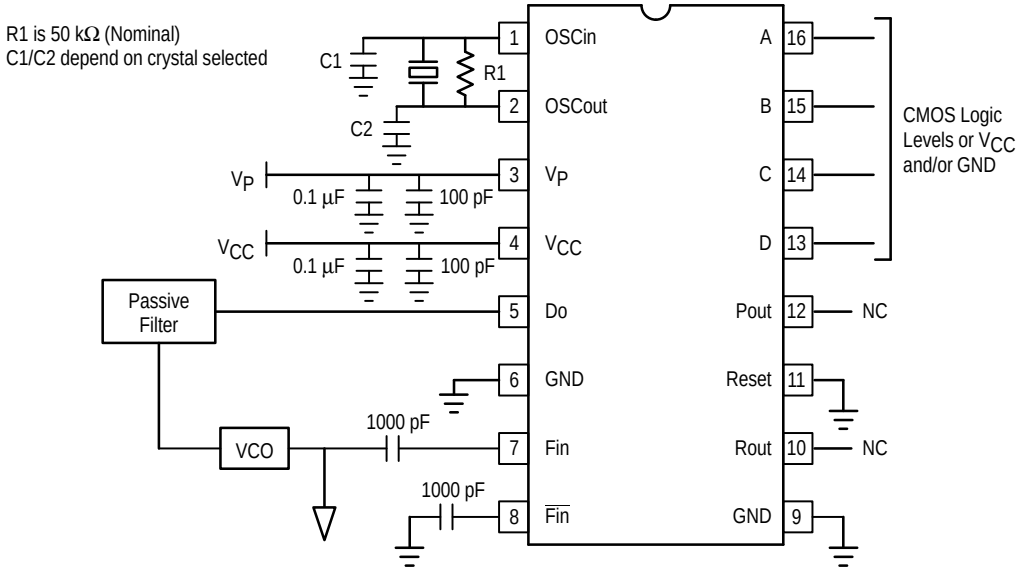
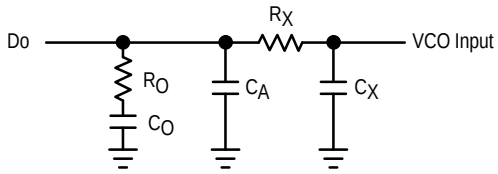


Figure 3. Typical Passive Loop Filter Topology



RECOMMENDED OPERATING CONDITIONS

Parameter	Symbol	Min	Max	Unit
Supply Range	V_{CC}	2.7	5.5	VDC
Maximum Supply Range	V_{CC}max	–	–6.0	VDC
Maximum Charge Pump Voltage	$V_P$max	–	V_{CC} to +6.0	VDC
Temperature Ambient	T_A	–40	85	°C
Storage Temperature	T_{STG}	–65	150	°C
Maximum Input Signal (Any Pin)	V_{in}max	–	V_{CC}+0.5 V	VDC

ELECTRICAL CHARACTERISTICS (V_{CC} = 2.7 to 5.5 V; V_P = V_{CC} to 6.0 V; T_A = –40 to +85°C, unless otherwise noted.)

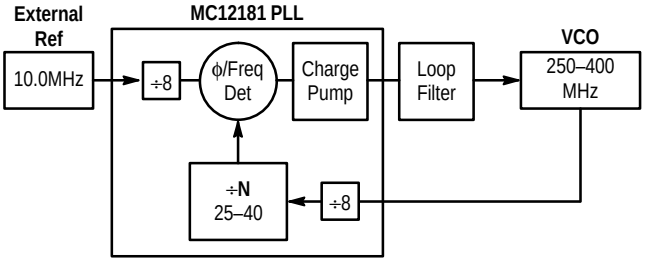
Characteristic	Symbol	Min	Typ	Max	Unit	Condition
Supply Current for V_{CC}	I_{CC}	–	4.0	5.5	mA	Note 1
Supply Current for V_P	I_P	–	0.25	0.5	mA	Note 1
Input Frequency Range	OSCin	5	–	25	MHz	Note 2
RF Input Frequency Range	Fin	125	–	1000	MHz	Note 3
Fin Input Sensitivity	Vin	100	–	1000	mVpp	Note 4
OSCin Input Sensitivity	V_{OSC}	500	–	2200	mVpp	Note 4
Output Source Current (Do)	I_{OH}	–2.8	–2.2	–2.0	mA	Note 5
		–2.4	–2.0	–1.6		Note 6
Output Sink Current (Do)	I_{OL}	2.0	2.4	2.8	mA	Note 5
		1.6	2.0	2.4		Note 6
Output Leakage Current (Do)	I_{OZ}	–	0.5	10	nA	V_{CC}=5.5; V_P = 6.0 V; VDo=0.5 to 5.5 V
Charge Pump Operating Volt	VDo	0.5	–	V_P–0.5	V	
Input HIGH Voltage Reset, A, B, C, D	V_{IH}	0.7 V_{CC}	–	–	V	
Input LOW Voltage Reset, A, B, C, D	V_{IL}	–	–	0.3 V_{CC}	V	
Input HIGH Current A, B, C, D	I_{IH}	–	–	+1	µA	
Reset		–	–	+100		
Input LOW Current A, B, C, D	I_{IL}	–100	–	–	µA	
Reset		–1	–	+1		
Output Amplitude (Pout, Rout)	Vout	250	400	–	mVpp	Note 7

NOTES: 1. V_{CC} and V_P = 5.5 V; Fin = 1.0 GHz; OSCin = 25 MHz; Do open.
2. Assumes C_1 and C_2 (Figure 2) limited to ≤30 pF each including stray capacitance in crystal mode, ac coupled input for external reference mode.
3. AC coupling, Fin measured with a 1000pF capacitor.
4. Signal ac coupling in input.
5. V_{CC} = 5.5 V; V_P = 6.0 V; V_{DO} = 3.0 V.
6. V_P = V_{CC} = 3.0 V; V_{DO} = 1.5 V.
7. Minimum resistor value of 25 kΩ to ground.

APPLICATIONS INFORMATION

The MC12181 is intended for applications where a fixed LO, or a limited number of local oscillator frequencies is required to be synthesized. The device acts as a x25 – 40 PLL. The 4–bit parallel interface allows 1 of 16 divide ratios to be selected. Internally there are fixed divide by 8 prescalers in the reference and programmable paths of the PLL. The MC12181 operates from 125 MHz to 1000 MHz which makes the part ideal for FCC Title 47; Part 15 applications in the 260 MHz to 470 MHz band and the 902 to 928 MHz Band. Figure 4 shows a typical block diagram of the application.

Figure 4. Typical Block Diagram of Complete PLL



As can be seen from the block diagram, with the addition of a VCO, a loop filter, and either an external oscillator or crystal, a complete PLL sub–system can be realized. Since most of the PLL functions are integrated into the 12181, the users focus is on the loop filter design and the crystal reference oscillator circuit.

Crystal Oscillator Design

The PLL is used to transfer the high stability characteristic of a low frequency reference source to the high frequency VCO within the PLL loop. To facilitate this, the device contains an input circuit which can be configured as a crystal oscillator or a buffer for accepting an external signal source.

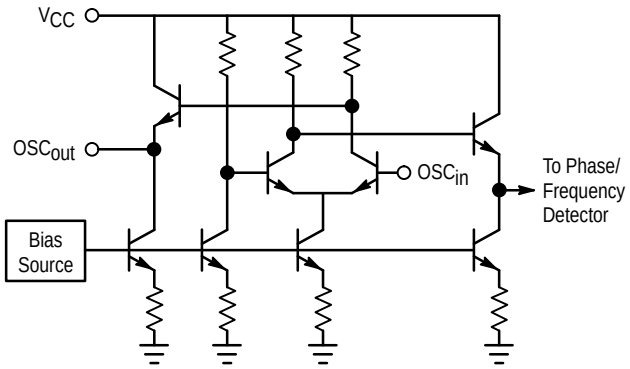
In the external reference mode, the reference source is ac–coupling into the OSCin input pin. The level of this signal should be between 500 – 2200 mVp–p. An external low noise reference should be used when it is desired to obtain the best close–in phase noise performance for the PLL. In addition the input reference amplitude should be close to the upper amplitude specification. This maximizes the slew rate of the input signal as it switches against the internal voltage reference.

In the crystal mode, an external parallel–resonant fundamental mode crystal should be connected between the OSCin and OSCout pins. This crystal must be between 5 and 25 MHz. External capacitors C1 and C2, as shown in Figure 2, are required to set the proper crystal load capacitance and oscillator frequency. The values of the capacitors are dependent on the crystal choosen and the input capacitance of the device as well as stray board capacitance.

Since the MC12181 is realized with an all–bipolar ECL style design, the internal oscillator circuitry is different from more traditional CMOS oscillator designs which realize the crystal oscillator with a modified inverter topology. These CMOS designs typically excite the crystal with a rail–to–rail signal which may overdrive the crystal resulting in damage or unstable operation. The MC12181 design does not exhibit this phenomena because the swing out of the OSCout pin is less than 600 mVp–p. This has the added advantage of minimizing EMI and switching noise which can be generated by rail–to–rail CMOS outputs. The OSCout output should not be used to drive other circuitry.

The oscillator buffer in the MC12181 is a single stage, high speed, differential input/output amplifier; it may be considered to be a form of the Pierce oscillator. A simplified circuit diagram is seen in Figure 5.

Figure 5. Simplified Crystal Oscillator/Buffer Circuit



OSC_{in} drives the base of one input of an NPN transistor differential pair. The non–inverting input of the differential pair is internally biased. OSC_{out} is the inverted input signal and is buffered by an emitter follower with a 70 μA pull–down current and has a voltage swing of about 600mVp–p. Open loop output impedance is approximately 425 Ω. The opposite side of the differential amplifier output is used internally to drive another buffer stage which drives the phase/frequency detector. With the 50 kΩ feedback resistor in place, OSC_{in} and OSC_{out} are biased to approximately 1.1 V below V_{CC}. The amplifier has a voltage gain of about 15dB and a bandwidth in excess of 150 MHz. Adherence to good RF design and layout techniques, including power supply pin decoupling, is strongly recommended.

A typical crystal oscillator application is shown in Figure 2. The crystal and the feedback resistor are connected directly between OSC_{in} and OSC_{out}, while the loading capacitors, C1 and C2, are connected between OSC_{in} and ground, and OSC_{out} and ground respectively. It is important to understand that as far as the crystal is concerned, the two loading capacitors are in series (albeit through ground). So when the crystal specification defines a specific loading capacitance, this refers to the total external (to the crystal) capacitance seen across its two pins.

This capacitance consists of the capacitance contributed by the amplifier (IC and packaging), layout capacitance, and the series combination of the two loading capacitors. This is illustrated in the equation below:

$$C_I = C_{AMP} + C_{STRAY} + \frac{C1 \times C2}{C1 + C2}$$

Provided the crystal and associated components are located immediately next to the IC, thus minimizing the stray capacitance, the combined value of C_{AMP} and C_{STRAY} is approximately 5pF. Note that the location of the OSC_{in} and OSC_{out} pins at the end of the package, facilitates placing the crystal, resistor and the C1 and C2 capacitors very close to the device. Usually, one of the capacitors is in parallel with an adjustable capacitor used to trim the frequency of oscillation.

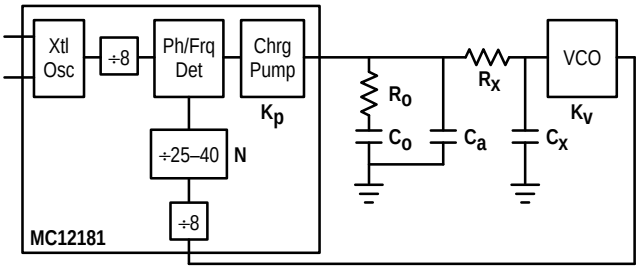
It is important that the total external (to the IC) capacitance seen by either OSC_{in} or OSC_{out}, be no greater than 30pF.

In operation, the crystal oscillator will start up with the application of power. If the crystal is in a can that is not grounded it is often possible to monitor the frequency of oscillation by connecting an oscilloscope probe to the can; this technique minimizes any disturbance to the circuit. If this is not possible, a high impedance, low capacitance, FET probe can be connected to either OSC_{in} or OSC_{out}. Signals typically seen at those points will be very nearly sinusoidal with amplitudes of roughly 300–600mVp–p. Some distortion is inevitable and has little bearing on the accuracy of the signal going to the phase detector.

Loop Filter Design

Because the device is designed for a non–frequency agile synthesizer (i.e., how fast it tunes is not critical) the loop filter design is very straight forward. The current output of the charge pump allows the loop filter to be realized without the need of any active components. The preferred topology for the filter is illustrated in Figure 6.

Figure 6. Loop Filter



The R_O/C_O components realize the primary loop filter. C_a is added to the loop filter to provide for reference sideband suppression. If additional suppression is needed, the R_X/C_X realizes an additional filter. In most applications, this will not be necessary. If all components are used, this results in a 4th order PLL, which makes analysis difficult. To simplify this, the loop design will be treated as a 2nd order loop (R_O/C_O) and additional guidelines are provided to minimize the influence of the other components. If more rigorous analysis is needed, mathematical/system simulation tools should be used.

Component	Guideline
C_a	$<0.1 \times C_O$
R_X	$>10 \times R_O$
C_X	$<0.1 \times C_O$

The focus of the design effort is to determine what the loop's natural frequency, ω_o, should be. This is determined by R_O, C_O, K_p, K_V, and N_t. Because K_p, K_V, and N_t are given, it is only necessary to calculate values for R_O and C_O. There are 3 considerations in selecting the loop bandwidth:

1) Maximum loop bandwidth for minimum tuning speed
2) Optimum loop bandwidth for best phase noise performance
3) Minimum loop bandwidth for greatest reference sideband suppression

Usually a compromise is struck between these 3 cases, however, for a fixed frequency application, minimizing the tuning speed is not a critical parameter.

To specify the loop bandwidth for optimal phase noise performance, an understanding of the sources of phase noise in the system and the effect of the loop filter on them is required. There are 3 major sources of phase noise in the phase–locked loop – the crystal reference, the VCO, and the loop contribution. The loop filter acts as a low–pass filter to the crystal reference and the loop contribution. The loop filter acts as a high–pass filter to the VCO with an in–band gain equal to unity. The loop contribution includes the PLL IC, as well as noise in the system; supply noise, switching noise, etc. For this example, a loop contribution of 15dB has been selected, which corresponds to data in Figure NO TAG.

The crystal reference and the VCO are characterized as high–order 1/f noise sources. Graphical analysis is used to determine the optimum loop bandwidth. It is necessary to have noise plots from the manufacturers of both devices. This method provides a straightforward approximation suitable for quickly estimating the optimal bandwidth. The loop contribution is characterized as white–noise or low–order 1/f noise given in the form of a noise factor which combines all the noise effects into a single value. The phase noise of the Crystal Reference is increased by the noise factor of the PLL IC and related circuitry. It is further increased by the total divide–by–N ratio of the loop. This is illustrated in Figure 7.

The point at which the VCO phase noise crosses the amplified phase noise of the Crystal Reference is the point of the optimum loop bandwidth. In the example of Figure 7, the optimum bandwidth is approximately 15 KHz.

Figure 7. Graphical Analysis of Optimum Bandwidth

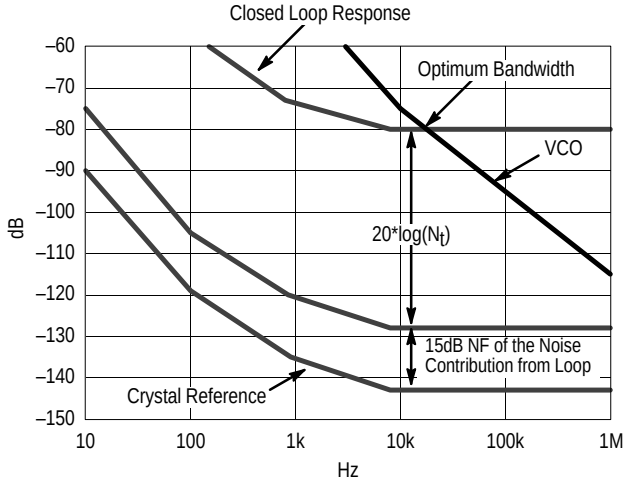
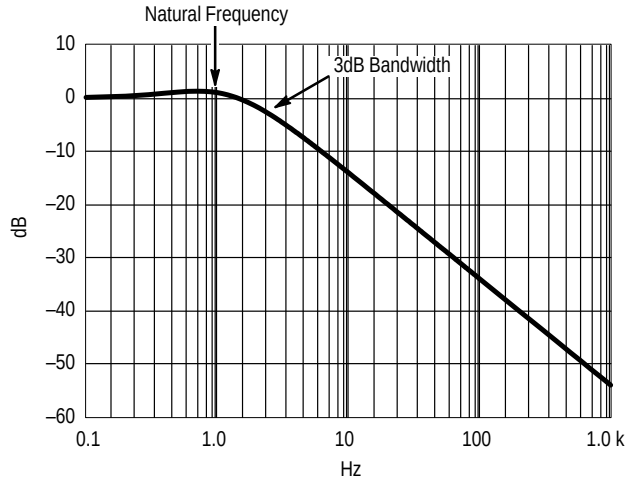


Figure 8. Closed Loop Frequency Response for ζ = 1



To simplify analysis further a damping factor of 1 will be selected. The normalized closed loop response is illustrated in Figure 8 where the loop bandwidth is 2.5 times the loop natural frequency (the loop natural frequency is the frequency at which the loop would oscillate if it were unstable). Therefore the optimum loop bandwidth is 15 kHz/2.5 or 6.0 kHz (37.7 krads) with a damping coefficient, $\zeta \approx 1$. T(s) is the transfer function of the loop filter.

$$T(s) = \frac{R_oC_os + 1}{\left(\frac{NC_o}{K_pK_v}\right)s^2 + R_oC_os + 1} = \frac{\left(\frac{2\zeta}{\omega_o}\right)s + 1}{\left(\frac{1}{\omega_o{}^2}\right)s^2 + \left(\frac{2\zeta}{\omega_o}\right)s + 1}$$

$$\left(\frac{NC_o}{K_pK_v}\right) = \left(\frac{1}{\omega_o{}^2}\right) \rightarrow \omega_o = \sqrt{\frac{K_pK_v}{NC_o}} \rightarrow \boxed{C_o = \left(\frac{K_pK_v}{N\omega_o{}^2}\right)}$$

$$R_oC_o = \left(\frac{2\zeta}{\omega_o}\right) \rightarrow \zeta = \left(\frac{\omega_oR_oC_o}{2}\right) \rightarrow \boxed{R_o = \left(\frac{2\zeta}{\omega_oC_o}\right)}$$

where N_t = Total PLL Divide Ratio — 8×N where (N = 25...40)
K_v = VCO Gain — Hz/V
K_p = Phase Detector/Charge Pump Gain — A
= ($|I_{OH}| + |I_{OL}|$) / 2

Technically, K_v and K_p should be expressed in Radian units [K_v (RAD/V), K_p (A/RAD)]. Since the component design equation contains the $K_v \times K_p$ term. the 2π cancels and the values can be epressed as above.

Figure 9. Design Equations for the 2nd Order System

In summary, follow the steps given below:

Step 1: Plot the phase noise of crystal reference and the VCO on the same graph.

Step 2: Increase the phase noise of the crystal reference by the noise contribution of the loop.

Step 3: Convert the divide–by–N to dB (20log 8 × N) and increase the phase noise of the crystal reference by that amount.

Step 4: The point at which the VCO phase noise crosses the amplified phase noise of the Crystal Reference is the point of the optimum loop bandwidth. This is approximately 15 kHz in Figure 7.

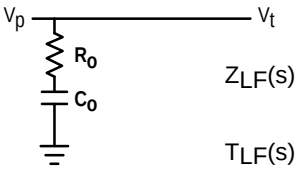
Step 5: Correlate this loop bandwidth to the loop natural frequency per Figure 8. In this case the 3.0 dB bandwidth for a damping coefficient of 1 is 2.5 times the loop's natural frequency. The relationship between the 3.0 dB loop bandwidth and the loop's "natural" frequency will vary for different values of ζ. Making use of the equations defined in Figure 9, a math tool or spread sheet is useful to select the values for R_o and C_o.

Appendix: Derivation of Loop Filter Transfer Function

The purpose of the loop filter is to convert the current from the phase detector to a tuning voltage for the VCO. The total transfer function is derived in two steps. Step 1 is to find the voltage generated by the impedance of the loop filter. Step 2 is to find the transfer function from the input of the loop filter to its output. The "voltage" times the "transfer function" is the overall transfer function of the loop filter. To use these equations in determining the overall transfer function of a PLL multiply the filter's impedance by the gain constant of the phase detector then multiply that by the filter's transfer function (Figure 10 contains the transfer function equations for 2nd, 3rd and 4th order PLL filters.)

Figure 10. Overall Transfer Function of the PLL

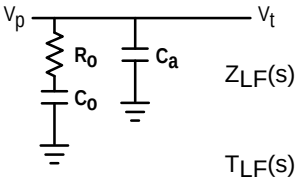
For the 2nd Order PLL:



$$Z_{LF}(s) = \frac{R_oC_os + 1}{C_os}$$

$$T_{LF}(s) = \frac{V_t(s)}{V_p(s)} = 1\ ,\ V_p(s) = K_p(s)Z_{LF}(s)$$

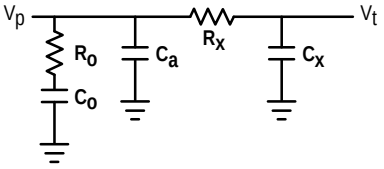
For the 3rd Order PLL:



$$Z_{LF}(s) = \frac{R_oC_os + 1}{C_oR_oC_as^2 + (C_o + C_a)s}$$

$$T_{LF}(s) = \frac{V_t(s)}{V_p(s)} = 1\ ,\ V_p(s) = K_p(s)Z_{LF}(s)$$

For the 4th Order PLL:



$$Z_{LF}(s) = \frac{(R_oC_os + 1)\ (R_xC_xs + 1)}{C_oR_oC_aR_xC_xs^3 + [\ (C_o + C_a)R_xC_x + C_oR_o(C_x + C_a)\ ]\ s^2 + (C_o + C_a + C_x)s}$$

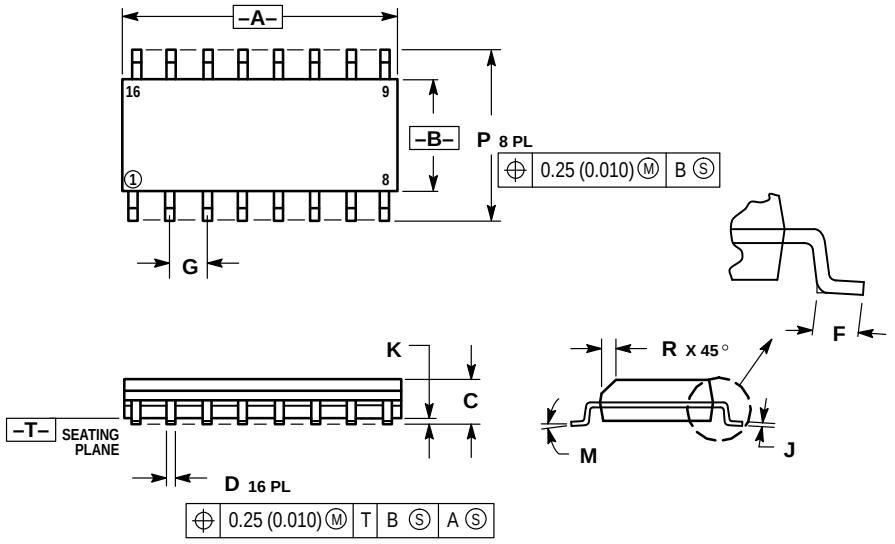
$$T_{LF}(s) = \frac{V_t(s)}{V_p(s)} = \frac{1}{(R_xC_xs + 1)}\ ,\ V_p(s) = K_p(s)Z_{LF}(s)$$

Figure 11. Typical Charge Pump Current versus Temperature
(V_{CC} = 5.5 V; V_P = 6.0 V)

FIGURES 11 THRU 17 COULD NOT BE PROCESSED FOR PDF FORMAT. FOR COMPLETE DOCUMENT WITH ALL IMAGES, PLEASE ORDER FROM MFAX OR LITERATURE DISTRIBUTION CENTER.

OUTLINE DIMENSIONS

D SUFFIX
PLASTIC PACKAGE
CASE 751B-05
(SO–16)
ISSUE J



NOTES:
1. DIMENSIONING AND TOLERANCING PER ANSI Y14.5M, 1982.
2. CONTROLLING DIMENSION: MILLIMETER.
3. DIMENSIONS A AND B DO NOT INCLUDE MOLD PROTRUSION.
4. MAXIMUM MOLD PROTRUSION 0.15 (0.006) PER SIDE.
5. DIMENSION D DOES NOT INCLUDE DAMBAR PROTRUSION. ALLOWABLE DAMBAR PROTRUSION SHALL BE 0.127 (0.005) TOTAL IN EXCESS OF THE D DIMENSION AT MAXIMUM MATERIAL CONDITION.

	MILLIMETERS		INCHES	
DIM	MIN	MAX	MIN	MAX
A	9.80	10.00	0.386	0.393
B	3.80	4.00	0.150	0.157
C	1.35	1.75	0.054	0.068
D	0.35	0.49	0.014	0.019
F	0.40	1.25	0.016	0.049
G	1.27 BSC		0.050 BSC	
J	0.19	0.25	0.008	0.009
K	0.10	0.25	0.004	0.009
M	0°	7°	0°	7°
P	5.80	6.20	0.229	0.244
R	0.25	0.50	0.010	0.019

How to reach us:
USA / EUROPE / Locations Not Listed: Motorola Literature Distribution;
P.O. Box 5405, Denver, Colorado 80217. 1–303–675–2140 or 1–800–441–2447

JAPAN: Nippon Motorola Ltd.: SPD, Strategic Planning Office, 141,
4–32–1 Nishi–Gotanda, Shagawa–ku, Tokyo, Japan. 03–5487–8488

Customer Focus Center: 1–800–521–6274

Mfax™: RMFAX0@email.sps.mot.com – TOUCHTONE 1–602–244–6609
Motorola Fax Back System – US & Canada ONLY 1–800–774–1848
– http://sps.motorola.com/mfax/

ASIA / PACIFIC: Motorola Semiconductors H.K. Ltd.; 8B Tai Ping Industrial Park,
51 Ting Kok Road, Tai Po, N.T., Hong Kong. 852–26629298

HOME PAGE: http://motorola.com/sps/