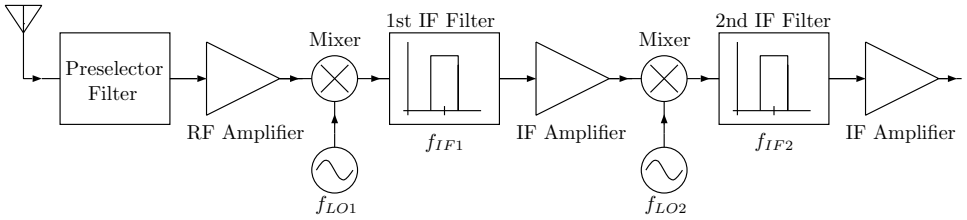
ECE 453
Wireless Communication Systems



Course Notes

Spring 2019

Steven J. Franke
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Contents

List of Tables

Chapter 1

Communication Signals and Systems

This chapter provides a brief overview of the basic nomenclature used to describe various parts of the radio frequency (RF) spectrum, followed by an overview of the dominant propagation modes that operate in various frequency ranges. Then, various methods that can be used to encode a message signal onto an RF carrier will be discussed.

1.1 The Radio Frequency Spectrum

In a communications system a message signal (e.g., analog video, analog audio, or a sequence of analog waveforms representing digital information) is encoded onto a carrier signal, which is transmitted and propagated to a receiver via one or more propagation modes. The process of encoding information onto a carrier is called *modulation* and the inverse process is called *demodulation*. In most cases the choice of carrier frequency is determined by the type of signal coverage that is desired, e.g., for world-wide coverage from a ground-based transmitter, one might choose a carrier frequency between 2 and 30 MHz (depending on the time of day and various other factors). In this frequency range signals can be propagated over long distances via the ionosphere "sky-wave" mode. On the other hand, for coverage of a particular geographic area from a satellite, the carrier frequency would need to be greater than 30 MHz (in order to penetrate the ionized regions of the upper atmosphere), and perhaps as high as 14 GHz. Of course, other considerations are also involved in the choice of carrier frequency. Because of the need to avoid interference between the various broadcast and communications services, frequency allocations for different purposes are strictly regulated. Within the United States the regulatory body is the Federal Communications Commission (FCC). The FCC conforms to international agreements established by the International Telecommunications Union (ITU). Periodically the frequency allocation scheme is revised in accordance with the needs of the various radio services and also to accommodate new technological developments. These revisions are agreed on by World Administrative Radio Conferences (WARCs) under the auspices of the ITU.

The basic principles, techniques and analysis methods that will be discussed herein are applicable over a broad range of frequencies. The terminology that is used to describe the various frequency bands is summarized in Table 1.1.

Table 1.1: Nomenclature for Frequency Bands.

Freq. Range	**Wavelength**	**Designation**	
30-3000 Hz	$10 - .1$ megameters	ELF	(Extremely Low Freq.)
3-30 kHz	$100 - 10$ kilometers	VLF	(Very Low Freq.)
30-300 kHZ	$10 - 1$ kilometers	LF	(Low Freq.)
300-3000 kHz	$1000 - 100$ meters	MF	(Medium Freq.)
3-30 MHz	$100 - 10$ meters	HF	(High Freq.)
30-300 MHz	$10 - 1$ meters	VHF	(Very High Freq.)
300-3000 MHz	$1 - 0.1$ meters	UHF	(Ultra High Freq.)
3-30 GHz	$10 - 1$ centimeters	SHF	(Super High Freq.)
30-300 GHz	$1 - 0.1$ centimeters	EHF	(Extremely High Freq.)
300-3000 GHz	$1 - .1$ millimeters	THF	(0.3-3.0 Terahertz)
3000-30000 GHz	$.1 - .01$ millimeters		(3.0-30.0 Terahertz)

Another nomenclature that is often employed, especially when referring to radar and satellite systems, is the so-called radar frequency band designations as defined by the IEEE and summarized in Table 1.2. The frequency allocations for common radio services are summarized in Table 1.3.

Table 1.2: Nomenclature for IEEE Radar Bands.

Frequency Range	**Designation**
0.23-1 GHz	P Band
1-2 GHz	L Band
2-4 GHz	S Band
4-8 GHz	C Band
8-12.5 GHz	X Band
12.5-18 GHz	Ku Band
18-26.5 GHz	K Band
26.5-40 GHz	Ka Band
40-75 GHz	V Band
75-110 GHz	W Band

Television broadcasts in the United States are allowed in several frequency bands within the range 54-698 MHz. Each assigned channel is 6.0 MHz wide. The frequency assignments and the associated channel numbers are summarized in Table 1.4.

1.2 Some Characteristics of Radio Propagation

Below approximately 10-20 MHz the ionosphere looks like a conducting "plate", or mirror, at a height of 80-400 km (see Figure 1.1). Signals with frequencies less than approximately 20 MHz are reflected back to earth when incident on the ionosphere from below. Signals with frequencies well above 20 MHz will pass right through the ionosphere. The upper frequency limit where signals will begin to penetrate the ionosphere depends on a number of factors, including the angle at which the signal is incident on the ionosphere, time of day, season,

Table 1.3: Frequency allocations for common radio services.

AM Broadcast band:	540-1700 kHz	10 kHz chan. BW
FM broadcast band:	88.0-108.0 MHz	200 kHz chan. BW
Aircraft band (civil aviation)	108.0-137.0 MHz	
Television	See Table 1.4.	
700 MHz cellular phone	698-806 MHz	
800 MHz cellular phone	824-849 MHz (mobile to base)	
	869-894 MHz (base to mobile)	
Global Positioning System (GPS)	1575.42 MHz (L1)	
	1227.60 MHz (L2)	
	1176.45 MHz (L5)	
PCS band phone	1850-1910 MHz (mobile to base)	
	1930-1990 MHz (base to mobile)	
WiFi, 802.11b/g	2400-2483.5 MHz	
Bluetooth	2400-2483.5 MHz	
WiFi, 802.11a	5150-5850 MHz	54 Mbps LAN

the amount of ionizing solar radiation, etc.

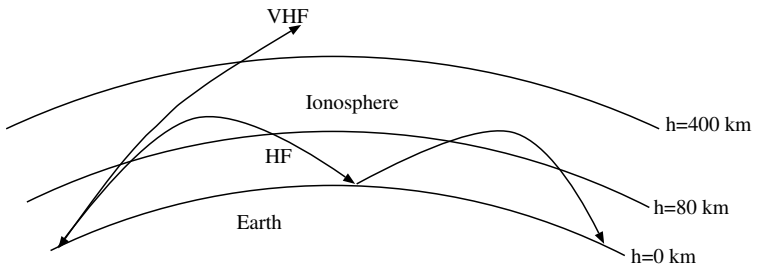


Figure 1.1: HF signals are refracted back toward the earth in the lower regions of the ionosphere. Depending on the frequency of the signal, passage through the ionosphere can result in effects ranging from virtually complete absorption of the signal to nearly lossless refraction of the signal. When absorption is low multiple earth-ionosphere-earth "hops" allow the signal to travel great distances. At HF, propagation completely around the world via many such "hops".

1.2.1 ELF through VLF

In this frequency range wavelengths are comparable to the separation between the earth's surface and the ionosphere; therefore the earth-ionosphere cavity looks like a waveguide. Because signals are confined to the earth-ionosphere waveguide, very long propagation distances are possible with relatively small signal attenuation. Frequencies at the low end of this range (ELF) are useful for communication with submarines, primarily due to the large skin depth at these frequencies, and hence relatively deep penetration of these signals in salt water.

Table 1.4: Television channel frequency assignments.

Channel Number	Frequency MHz	Channel Number	Frequency MHz
2	54-60	27	548-554
3	60-66	28	554-560
4	66-72	29	560-566
5	76-82	30	566-572
6	82-88	31	572-578
7	174-180	32	578-584
8	180-186	33	584-590
9	186-192	34	590-596
10	192-198	35	596-602
11	198-204	36	602-608
12	204-210	37	608-614
13	210-216	38	614-620
		39	620-626
14	470-476	40	626-632
15	476-482	41	632-638
16	482-488	42	638-644
17	488-494	43	644-650
18	494-500	44	650-656
19	500-506	45	656-662
20	506-512	46	662-668
21	512-518	47	668-674
22	518-524	48	674-680
23	524-530	49	680-686
24	530-536	50	686-692
25	536-542	51	692-698
26	542-548		

1.2.2 LF and MF

A frequency in the LF range (60 kHz) is employed for the time-signal broadcasts used by many of the clocks and wristwatches that feature automatic setting and "atomic-standard" accuracy. In the continental US, these signals originate in Colorado, from radio station WWVB. The "AM" broadcast band (540-1700 kHz) is in the MF range. During daytime the ionospherically reflected MF signals are strongly attenuated by absorption in the lower levels of the ionosphere. Therefore daytime propagation is primarily via the "ground wave" which is a term used to describe the component of the signal that travels along the earth's surface. Propagation somewhat beyond the line of sight is possible in daytime because the ground wave follows the earth's curvature. At night there is little ionospheric absorption and the "sky wave" can travel great distances via multiple hops between the reflecting ionosphere and the ground.

1.2.3 HF (Shortwave)

HF is similar to MF, but absorption during daytime is much less severe. Local reception is via ground wave and very long distance reception is via sky wave. For long distances, multiple sky wave paths are almost always present, resulting in signal fading as the multiple paths interfere constructively and destructively.

1.2.4 VHF (FM Broadcast, Television Channels 2-13)

Under most conditions the ionosphere has no effect on the signal, i.e. signals pass through the ionosphere. Propagation is usually line-of-sight via a direct path and, possibly, one or more reflected paths due to the ground and conducting objects located between the transmitter and receiver. Occasionally, relatively small patches of high electron density form in the E-region of the ionosphere (100-150 km in altitude) with densities sufficiently high so as to cause VHF signals to be reflected from the patches. This phenomenon is called "sporadic E". Figure 1.2 shows how these patches can reflect signals in the lower VHF range and cause long-distance reception (and possible interference).

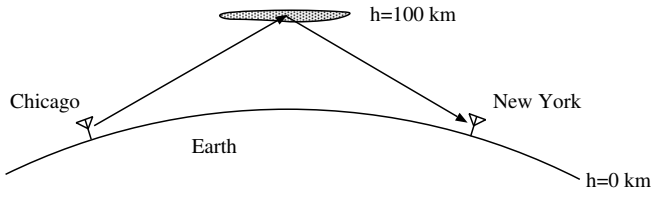


Figure 1.2: VHF signals can be reflected by thin and localized ionized layers in the E-region of the ionosphere at approximately 100 km altitude. While the occurrence of such layers is quite common in the summer and winter, the exact location and time of occurrence cannot be predicted, hence the name Sporadic-E layers.

1.2.5 UHF (Television Channels 14-51, cellular telephones, WiFi)

Similar to VHF but without significant sporadic-E propagation. At both VHF and UHF over-the-horizon propagation can occur due to refraction and tropospheric scatter. The term "tropospheric scatter" refers to weak scatter from turbulence-induced irregularities in the refractive index of the air within the troposphere. The troposphere includes the lower 10 km of the atmosphere.

Figure 1.3 illustrates refraction - slight ray bending due to altitude dependence of atmospheric humidity and temperature. Figure 1.4 illustrates tropospheric-scatter ("troposcatter") - scattering of energy off of atmospheric turbulence. Refraction and tropo-scatter are omni-present, but the amount of refractive bending and the strength of tropospheric scatter are variable and they depend on meteorological conditions in the troposphere between the transmitter and receiver sites.

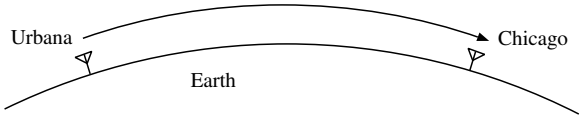


Figure 1.3: Waves propagating nearly horizontally in the troposphere tend to be refracted such that the propagation path follows the curvature of the earth. The refraction occurs because temperature and humidity naturally decrease with altitude, causing the refractive index to decreases with height from a value slightly larger than 1 (e.g. n$\simeq$1.0003) at the earth's surface to 1 at very high altitudes.

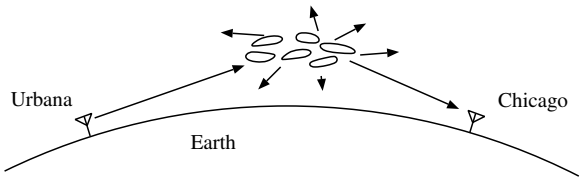


Figure 1.4: Turbulence in the atmosphere causes slight variations in refractive index which scatter radio signals in all directions.

1.3 Power Transfer via Free Space Propagation Links

Consider a free space propagation channel consisting of a transmitting antenna and a receiving antenna separated by distance R. The signal of interest has free space wavelength denoted by λ. The following discussion assumes that the distance between the antennas is large enough so that a signal radiated by either antenna can be approximated as having a planar constant phase surface in the vicinity of the other antenna. If the largest dimension of either antenna is denoted by D, the planar phase front approximation will be appropriate as long as $R \geq 2D^2/\lambda$.

If power P_t is delivered to the transmitting antenna, some of that power will be dissipated as heat in the antenna (P_d) and the rest of the power will be radiated into space (P_{rad}) so

$P_t = P_{rad} + P_d$. Denote the transmitting antenna power gain in the direction of the receiver by G_t. This gain is defined to be the ratio of the actual power density at the receiver location to the power density that would have resulted if the total power $P_t = P_{rad} + P_d$ had been radiated isotropically (uniformly) in all directions. This gain is therefore the gain relative to a hypothetical lossless, isotropic radiator and it includes the effects of power loss due to dissipation in the antenna and any concentration or dimunition of power density due to the coherent reinforcement or cancellation of fields radiated by different differential elements of the antenna structure. When expressed in decibels, the antenna gain is given the units "dBi" to remind us that the reference is the isotropic radiator. It is important to remember that G_t depends on direction, i.e. it is a function of two angles (θ, ϕ). The incident power density W_i at the receiving antenna can now be written as:

$$W_i = \frac{P_t G_t}{4\pi R^2}. \tag{1.1}$$

Assume that the receiving antenna's polarization is matched to the polarization of the incident signal and define an effective capture area A_r for the receiving antenna so that the power available from the receiving antenna's terminals, denoted by P_{avr}, can be written as the product of the incident power density and the capture area:

$$P_{avr} = W_i A_r = \frac{P_t G_t}{4\pi R^2} A_r. \tag{1.2}$$

Note that the available power, P_{avr}, is the power that the antenna would deliver to a conjugately-matched load. This is the basic communications link formula.

If the communications link is reciprocal (which will be true if the propagation medium is free space), then the roles of the transmitter and receiver can be reversed without changing the power transfer between the antennas. In other words, if the same total power P_t is delivered to the receiving antenna, the power available from the transmitting antenna will be P_{avr}, i.e.:

$$P_{avr} = \frac{P_t G_r}{4\pi R^2} A_t \tag{1.3}$$

where G_r is the gain of the receiving antenna in the direction to the transmitter and A_t is the effective capture area of the transmitting antenna. It follows that

$$G_t A_r = G_r A_t \tag{1.4}$$

or

$$G_t / A_t = G_r / A_r \tag{1.5}$$

This result was obtained without making any specific assumptions about the transmitting and receiving antennas, so the ratio G/A must be a universal quantity, independent of antenna details. This is a consequence of reciprocity. Generally speaking, propagation links will be reciprocal as long as the signal does not propagate through a medium (such as the ionosphere) that changes (e.g. rotates) the polarization of the electromagnetic fields.

Since the ratio G/A is a universal quantity, we can evaluate it once using results for any particular antenna, and the result must hold for any other antenna. This allows us to pick a simple antenna that allows for an easy calculation of G and A. The simple cases that are usually considered are those of a Hertzian dipole or a large uniformly illuminated aperture. In any case, the ratio G/A is found to be (e.g. see [4]):

$$\frac{G}{A} = \frac{4\pi}{\lambda^2} \tag{1.6}$$

so the relationship between capture area and antenna gain is

$$A = \frac{\lambda^2 G}{4\pi} \tag{1.7}$$

This relationship allows us to write the equation that governs power transfer in a radio link in terms of gains or capture area only, i.e.

$$P_{avr} = P_t \frac{G_t G_r \lambda^2}{16\pi^2 R^2} \tag{1.8}$$

or

$$P_{avr} = P_t \frac{A_t A_r}{\lambda^2 R^2} \tag{1.9}$$

The last version is known as the Friis transmission formula [4]. Friis liked this form of the link equation because it does not contain any numerical coefficients, a feature that he claimed would make it easier to remember.

1.3.1 Link example

The Friis transmission formula can be combined with knowledge of the operating system temperature (defined and discussed in Chapter 10) and required SNR to predict the performance of a communications link. The signal to noise ratio in an antenna-receiver system can be written as the ratio of available signal and noise powers at the antenna terminals, after referring all receiver noise to the antenna by assigning an operating system temperature $T_{op} = T_A + T_e$ to the antenna terminals. Here, T_A is the effective antenna temperature and T_e is the effective input noise temperature of the receiver:

$$SNR = \frac{P_{avr}}{kT_{op}B_n} = P_t \frac{G_t G_r \lambda^2}{16\pi^2 R^2} \frac{1}{kT_{op}B_n}$$

If the minimum SNR required for communications is denoted by SNR_{min}, then the maximum distance between the antennas is:

$$R_{max} = \sqrt{\frac{P_t G_t G_r \lambda^2}{SNR_{min}(16\pi^2 k T_{op} B_n)}}$$

For example, suppose that detection and/or demodulation of a particular signal requires an SNR of 2 (3 dB). We'll assume that the receiver filters define a noise bandwidth B_n equal to 10 kHz, that the transmitter power is (P_t) is 1 Watt, and the receiver and transmitting antennas are both lossless dipole antennas with length of one half wavelength, with the dipoles oriented in space so as to maximize the received signal power. The maximum gain of the lossless (100% efficient) half-wave dipole antenna is $G_t = G_r = 1.64$, or $10\log(1.64) = 2.14$ dBi.

Suppose that the link operates at a frequency of 300 MHz ($\lambda = 1$ m) and that the operating system temperature $T_{op} = 1000$ K (typical for an operating frequency in the VHF part of the spectrum). The maximum distance between the the two antennas in this case

is 2.4×10^7 m, or 24,000 km. This tells us that the power required to communicate over a distance of 24,000 km is comparable to that required to power a small flashlight!

Notice that $R_{max} \sim \lambda$ so that increasing (decreasing) the frequency will decrease (increase) R_{max}. At the same time, the size of both the transmitting and receiving antennas would decrease (increase) since they are assumed to be half-wavelength dipoles in this example. It is worthwhile to think about what would happen if, instead of keeping the antenna length fixed in terms of wavelengths, the absolute size of the transmitting and receiving antennas was held constant while the frequency is varied. This is left as a thought experiment to be carried out by the reader.

1.4 Review of Fourier Transforms and Spectra

Let $f(t)$ be a real-valued time-signal. The Fourier transform of $f(t)$ will be denoted by $F(\omega)$ where

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \tag{1.10}$$

The time-signal can be recovered from its Fourier transform using

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{+j\omega t} d\omega \tag{1.11}$$

The reader is assumed to be familiar with the general properties of Fourier transforms and transform pairs, e.g., the scaling and shifting properties. Time signals of interest are assumed to be absolutely integrable or to be periodic functions. In either case, the Fourier transform and its inverse will exist.

The magnitude of the Fourier transform, $|F(\omega)|$ is called the amplitude spectrum. For signals with finite energy, the square of the magnitude, $|F(\omega)|^2$ is called the energy spectrum since, by Parseval's theorem, the total energy, E, in the signal is

$$E = \int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega. \tag{1.12}$$

When the time-domain signal is real, the negative frequency part of the Fourier transform is related to the positive frequency part through

$$F(-\omega) = F^*(\omega) \tag{1.13}$$

where the * denotes complex conjugate. Thus if the positive part of the frequency spectrum is known, the negative frequency part can be obtained, since the amplitude of the spectrum is always symmetric about the frequency origin ($|F(-\omega)| = |F(\omega)|$) and the phase of the spectrum is anti-symmetric ($arg[F(-\omega)] = -arg[F(\omega)]$). For real signals, it is sometimes convenient to plot only the positive frequency part of the amplitude or energy spectrum.

1.5 Message signals

We will call the signal that is to be transmitted over the communications channel the *message signal*, denoted by $m(t)$. In an analog communications system the message signal could be the audio signal from a microphone. For example, Figure 1.5 (top) shows 0.8 seconds of

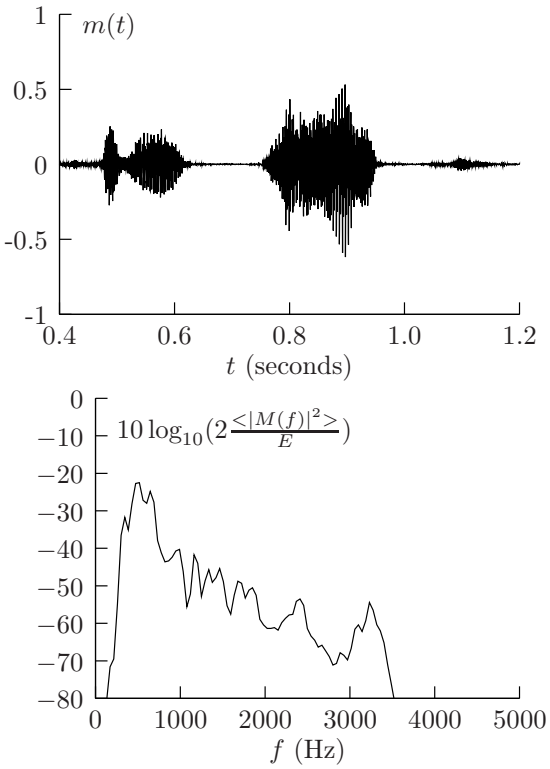


Figure 1.5: Top: message signal consisting of 0.8 seconds of audio from a radio talk show. Bottom: normalized average energy spectrum (in dB) of a 15-second segment of audio from the same talk show. The average energy spectrum was calculated by segmenting the original time series, sampled at 44.1 ksamples/s, into 512 point segments. The discrete Fourier transform (DFT) was calculated for each segment, and the squared magnitude of each DFT was averaged.

audio from a radio talk show. Figure 1.5 (bottom) shows the average energy spectrum calculated using an audio sample of duration 15 seconds. This audio sample contained male and female speakers. Notice that the energy is contained within the frequency range 200-3500 Hz. We shall use the symbol W to refer to the bandwidth of the message signal. The message signal bandwidth is defined such that the frequency interval $[-W, W]$ contains all (or, essentially all) of the energy in the signal. For the signal shown in Figure 1.5, W is approximately 3.5 kHz.

In a digital communication system the message signal will typically be formed from a superposition of pulses with amplitudes chosen to represent a sequence of information bits. In this case $m(t)$ can be written

$$m(t) = \sum_n a_n p(t - nT) \tag{1.14}$$

where $p(t)$ is a pulse function chosen to control the shape of the amplitude spectrum of $m(t)$, $\{a_n\}$ represents the sequence of information bits and is a sequence of numbers chosen from a finite set of possible values, and T^{-1} is the signaling rate — the rate at which pulses are transmitted. In the case of binary data, the pulse amplitudes would typically be chosen from the antipodal set $\{-1, 1\}$, so that $a_n = \pm 1$. In this case each information bit results in one transmitted pulse so that 1 bit of information is transmitted each T seconds. More generally, a_n could take on any of M discrete values, in which case we have an $M - ary$ data sequence and $\log_2 M$ bits are transmitted each T seconds. In general, the number of information bits transmitted per second will be the signaling rate (T^{-1}) times the number of bits represented by each pulse that is transmitted ($\log_2 M$).

Pulse-shapes that are commonly used to generate the message signal include the rectangular pulse (Figure 1.6a):

$$p(t) = \begin{cases} 1 & , \quad -\frac{1}{2} \leq \frac{t}{T} < \frac{1}{2} \\ 0 & , \quad \text{elsewhere} \end{cases},$$

with Fourier transform

$$P(f) = \frac{\sin \pi f T}{\pi f T}.$$

The half-cosine pulse (Figure 1.6b):

$$p(t) = \begin{cases} \cos \frac{\pi t}{T} & , \quad -\frac{1}{2} \leq \frac{t}{T} < \frac{1}{2} \\ 0 & , \quad \text{elsewhere} \end{cases},$$

with Fourier transform

$$P(f) = \frac{\sin \pi T(\frac{1}{2} + f)}{\pi T(\frac{1}{2} + f)} + \frac{\sin \pi T(\frac{1}{2} - f)}{\pi T(\frac{1}{2} - f)}$$

has a Fourier spectrum with a wider central lobe than that of the rectangular pulse, but the sidelobes are significantly smaller than those of the rectangular pulse. The rectangular and half-cosine pulses have infinite bandwidth, however the half-cosine pulse has a more compact spectrum because the amplitude of the half-cosine pulse's spectrum falls off much more rapidly with increasing frequency.

A very useful family of pulses with finite bandwidth is obtained by defining the pulse such that its Fourier transform transitions from a constant (flat) central region to zero through a

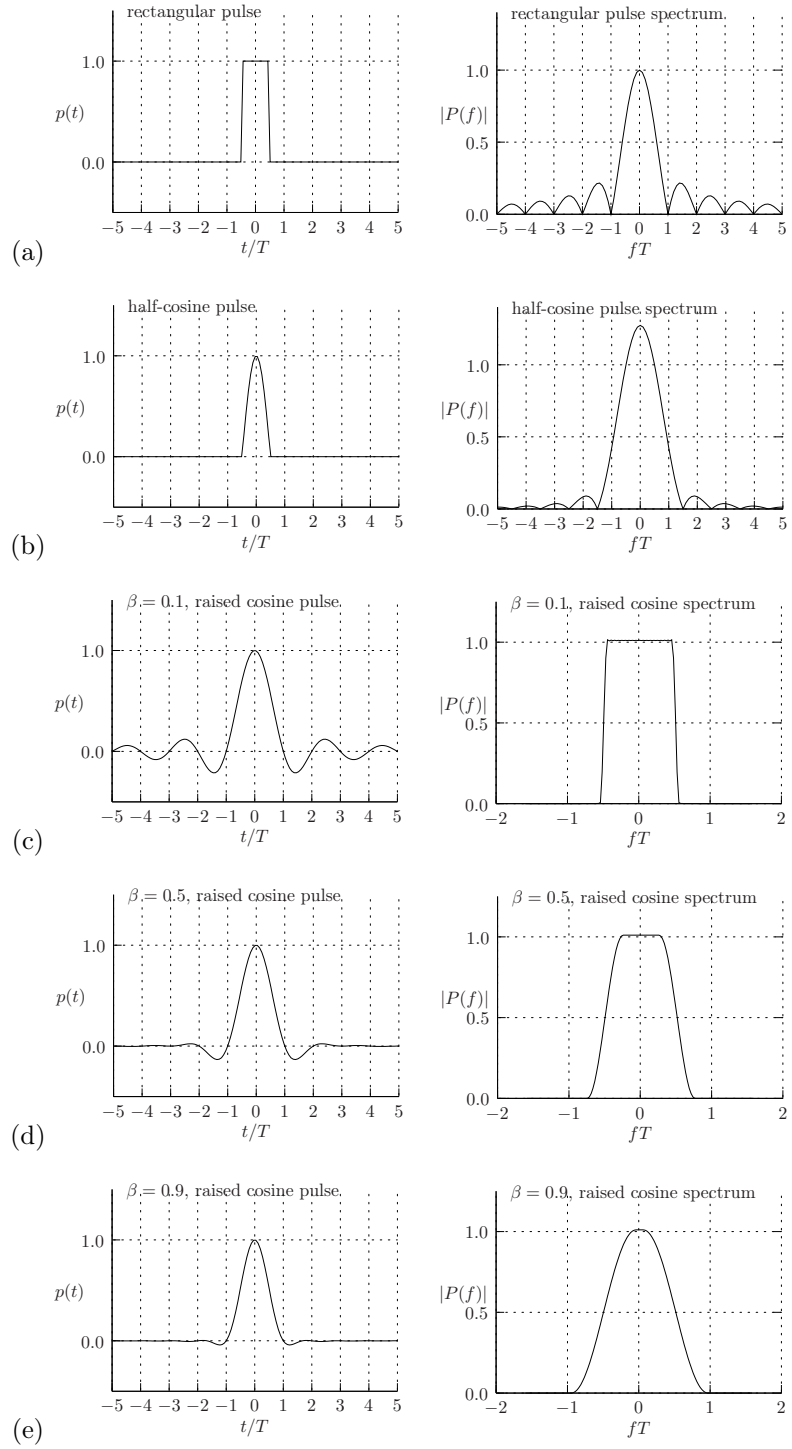


Figure 1.6: Pulse shapes suitable for use in digital communications systems. The time-domain pulse waveform is shown on the left and the magnitude of the Fourier Transform of the pulse is shown to the right. (a) rectangular pulse, (b) half-cosine pulse, (c)-(e) pulses with raised-cosine spectrum with excess bandwidths of (c) 10% ($\beta = 0.1$), (d) 50% ($\beta = 0.5$) and (e) 90% ($\beta = 0.9$).

smooth transition region having a raised cosine shape. These pulses are called *raised-cosine pulses* (Figure 1.6(c-e)). The Fourier transform of the raised-cosine pulse is

$$P(f) = \begin{cases} T & |fT| \leq \frac{1}{2}(1-\beta) \\ \frac{T}{2}\{1+\cos[\frac{\pi}{\beta}(|fT| - \frac{1}{2}(1-\beta)]\} & \frac{1}{2}(1-\beta) < |fT| \leq \frac{1}{2}(1+\beta) \\ 0 & |fT| > \frac{1}{2}(1+\beta) \end{cases}, \quad (1.15)$$

where $0 \leq \beta \leq 1$. The dimensionless parameter β controls the width of the raised-cosine transition region. The time-domain pulse shape of the raised-cosine pulse is

$$p(t) = \frac{\sin(\pi t/T)}{\pi t/T} \frac{\cos(\pi \beta t/T)}{1-(2\beta t/T)^2}. \quad (1.16)$$

The bandwidth of these pulses is

$$W = \frac{1}{2T}(1+\beta). \quad (1.17)$$

The dimensionless parameter β is called the fractional *excess bandwidth* and it is often expressed as a percentage. The minimum possible bandwidth results when $\beta = 0$ which results in $W = \frac{1}{2T}$. In this case, the spectrum becomes rectangular and the pulse shape is a sinc function with relatively large sidelobes that extend over many signaling intervals on either side of the center of the pulse. For $0 < \beta \leq 1$, the amplitude spectrum exhibits a gradual transition to zero. As β increases, the sidelobes are increasingly damped causing the time-domain pulse's sidelobes to have significant amplitudes over fewer signaling intervals.

It is important to point out that the spectrum of the chosen pulse will determine the shape of the spectrum of the message signal $m(t)$ and hence the bandwidth occupied by the message signal. Let us assume that the summation in equation 1.14 is finite and consists of N terms. The resulting $m(t)$ has finite energy. The Fourier transform of $m(t)$ is

$$M(f) = \mathcal{F}[m(t)] = \sum_{n=1}^{N} a_n \mathcal{F}[p(t-nT)] = P(f) \sum_{n=1}^{N} a_n \mathrm{e}^{-j2\pi fnT}. \quad (1.18)$$

The energy spectrum is

$$|M(f)|^2 = |P(f)|^2 |\sum_{n=1}^{N} a_n \mathrm{e}^{-j2\pi nfT}|^2. \quad (1.19)$$

For a specific data sequence $\{a_n\}$ the second term on the right-hand side will vary in an irregular manner as a function of frequency. Nevertheless, the second term has a well-defined average value. If $a_n = \pm 1$ the average value is equal to the length of the data sequence, N. Therefore, if the energy spectra of many message signals resulting from different random data sequences of length N are averaged, the average of the second term will be close to N. Using brackets $<>$ to denote an average over a large number of message signals,

$$< |M(f)|^2 > \simeq N|P(f)|^2, \quad (1.20)$$

which means that the energy spectrum of the pulse determines the average energy spectrum of the message signal. This is illustrated in Figure 1.7 which shows a portion of a message

signal generated using raised-cosine pulses with $\beta = 0.5$ and the average energy spectrum calculated by dividing the long message signal into shorter segments and averaging the energy spectra of each segment. Note that the shape of the average energy spectrum is determined by the raised-cosine spectrum of the individual pulses.

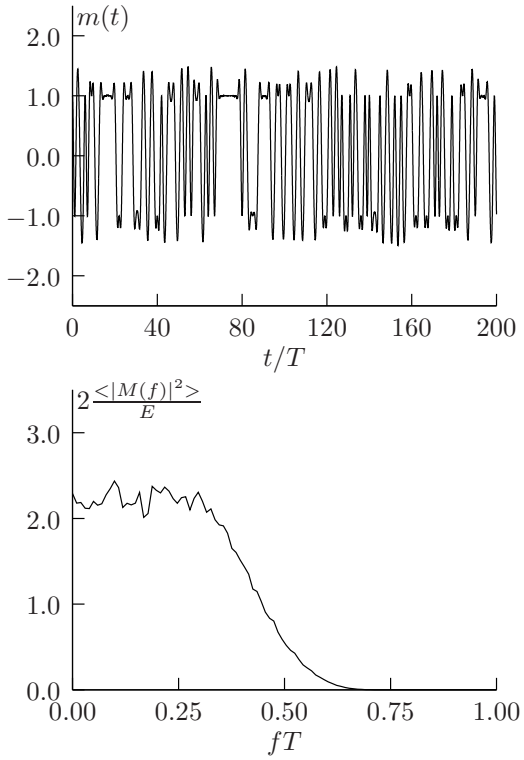


Figure 1.7: Top: A message signal produced using a random data sequence and raised-cosine pulses with $\beta = 0.5$. Bottom: Normalized average energy spectrum of $m(t)$ computed numerically using the following procedure: a message signal of length $32768T$ was generated using a random sequence of bits. The long message signal was divided into short segments of length approximately $100T$. The energy spectrum of each short segment was computed, and all spectra were averaged. The average spectrum was normalized so that the area under the spectrum is 1.0. Note that the shape of the energy spectrum exhibits the raised-cosine shape of the pulses. Since $\beta = 0.5$, the bandwidth of the spectrum is $WT = \frac{1}{2}(1 + 0.5) = 0.75$.

The raised-cosine pulses are *Nyquist pulses*, having the property that they are zero at non-zero multiples of the signaling interval T. Because of this property, such pulses can be superposed as in equation 1.14 without causing any inter-pulse (or intersymbol) interference at times corresponding to integer multiples of the signaling interval ($t = nT$) because only the pulse centered at $t = nT$ contributes to $m(t)$ at those times. Refer ahead to the upper plot in Figure 1.8 to see a sample message signal generated from a binary pulse amplitude sequence ($a_n = \pm 1$) and raised-cosine pulses with $\beta = 0.1$. The binary message can be read off of the upper plot by taking samples at integer values of t/T; in this case, the message

was the sequence

$$\{1, -1, 1, 1, 1, -1, 1, -1, -1, 1, 1, 1, 1, -1, 1, -1, 1, -1, -1, 1, -1\}.$$

In the following sections the basic types of analog modulation will be discussed. They fall into two general categories: (1) Linear and (2) Nonlinear (or angle) modulation schemes. Each category can be subdivided into special cases:

1. Linear modulation:
   (a) DSB-SC - Double Sideband-Suppressed Carrier
   (b) DSB with carrier
   (c) SSB - Single Sideband
   (d) VSB - Vestigial Sideband
2. Angle modulation (nonlinear modulation):
   (a) FM - Frequency Modulation
   (b) PM - Phase Modulation

1.6 Linear Modulation

Throughout all of our discussions we will assume that the message signal, $m(t)$, is a real function of time and that the time-average value of $m(t)$ is zero ($m(t)$ has no DC component). The linear modulation schemes can be represented by

$$s(t) = A(t)\cos(\omega_c t + \theta) \tag{1.21}$$

where $s(t)$ is called the modulated carrier signal, $A(t) = Am(t) + B$, ω_c is the carrier frequency, and θ is a constant phase angle. The magnitude of the instantaneous amplitude, $|A(t)|$ is called the *envelope* of $s(t)$.

The various types of linear modulation come about from different choices for B (e.g. $B = 0$ or $B > \max|m(t)|$) or, in the case of single-sideband (SSB), from superposing two or more linearly-modulated signals, e.g., $A_1(t)\cos(\omega_c t + \theta_1) + A_2(t)\cos(\omega_c t + \theta_2)$ where $A_1(t)$ and $A_2(t)$ are are obtained by passing $m(t)$ through different filters. The linear modulation schemes are "linear" in the sense that if $m_1(t) \to s_1(t)$ and $m_2(t) \to s_2(t)$, then $m_1(t) + m_2(t) \to s_1(t) + s_2(t)$[1].

1.6.1 Modulation Theorem

Before considering the various linear modulation schemes it will be useful to derive a fundamental relationship between the Fourier transforms of $s(t)$ and $A(t)$. Denote the Fourier transforms of $s(t)$ and $A(t)$ by $S(\omega)$ and $A(\omega)$, respectively; i.e.,

$$S(\omega) = \int_{-\infty}^{\infty} s(t)e^{-j\omega t}dt \tag{1.22}$$

[1]Strictly speaking, this linearity property applies only when $B = 0$.

$$A(\omega) = \int_{-\infty}^{\infty} A(t)e^{-j\omega t}dt. \tag{1.23}$$

Substituting Equation 1.21 into Equation 1.22 we obtain:

$$S(\omega) = \int_{-\infty}^{\infty} A(t)\cos(\omega_c t + \theta)e^{-j\omega t}dt. \tag{1.24}$$

Using the identity $\cos(\theta) = \frac{1}{2}\left[e^{j\theta} + e^{-j\theta}\right]$, Equation 1.24 becomes

$$\begin{aligned} S(\omega) &= \int_{-\infty}^{\infty} A(t)\frac{1}{2}\left[e^{j(\omega_c t+\theta)} + e^{-j(\omega_c t+\theta)}\right]e^{-j\omega t}dt \qquad (1.25)\\ &= \frac{1}{2}\,e^{j\theta}\int_{-\infty}^{\infty} A(t)e^{-j(\omega-\omega_c)t}dt + \frac{1}{2}e^{-j\theta}\int_{-\infty}^{\infty} A(t)e^{-j(\omega+\omega_c)t}dt \end{aligned}$$

which leads to the final result:

$$S(\omega) = \frac{1}{2}e^{j\theta}A(\omega-\omega_c) + \frac{1}{2}\,e^{-j\theta}A(\omega+\omega_c). \tag{1.26}$$

This important relationship between the spectrum of $A(t)$ and the spectrum of the modulated carrier $s(t)$ is known as the *modulation theorem.* It states that the spectrum of $s(t)$ can be obtained by superposing two copies of the spectrum of $A(t)$ that have been displaced by $+\omega_c$ and $-\omega_c$ on the frequency axis.

1.6.2 Double-sideband suppressed-carrier (DSB-SC) Modulation and Demodulation

DSB-SC represents the simplest possible mapping between the message signal and $A(t)$. Here we take $A(t) = Am(t)$ where A is a constant, i.e.,

$$s(t) = Am(t)\cos(\omega_c t + \theta) \tag{1.27}$$

This corresponds to multiplying the carrier by the message signal as illustrated in Figure 1.8, where the message signal is shown in the upper plot and the modulated carrier is shown below. For this example, the message signal is generated from a binary pulse amplitude sequence $\{a_n\}$ using raised-cosine pulses with $\beta = 0.1$. The binary message can be read off of the upper plot by taking samples at integer values of t/T; in this case, the message was the sequence

$$\{1, -1, 1, 1, 1, -1, 1, -1, -1, 1, 1, 1, 1, -1, 1, -1, 1, -1, -1, 1, -1\}.$$

The DSB-SC signal, $s(t)$, shown in the lower panel has been normalized to have a mean-square value of 1.0. Thus, if this signal represents the voltage developed across a $1\,\Omega$ resistor, the average power dissipated in the resistor would be 1 Watt.

Using the modulation theorem, it is a simple matter to relate the Fourier spectra of $m(t)$ and $s(t)$. Suppose that the two-sided spectrum of $m(t)$ is as shown in Figure 1.9 where $m(t)$ is assumed to be band-limited to W Hz. For simplicity, we plot only the magnitude of the spectrum. Then, if the carrier frequency is larger than $W(f_c > W)$, the spectrum of $s(t)$ will look like Figure 1.10. If $f_c < W$, the picture would be different, since the two components of the spectrum would overlap in a region centered on $f = 0$. To avoid such

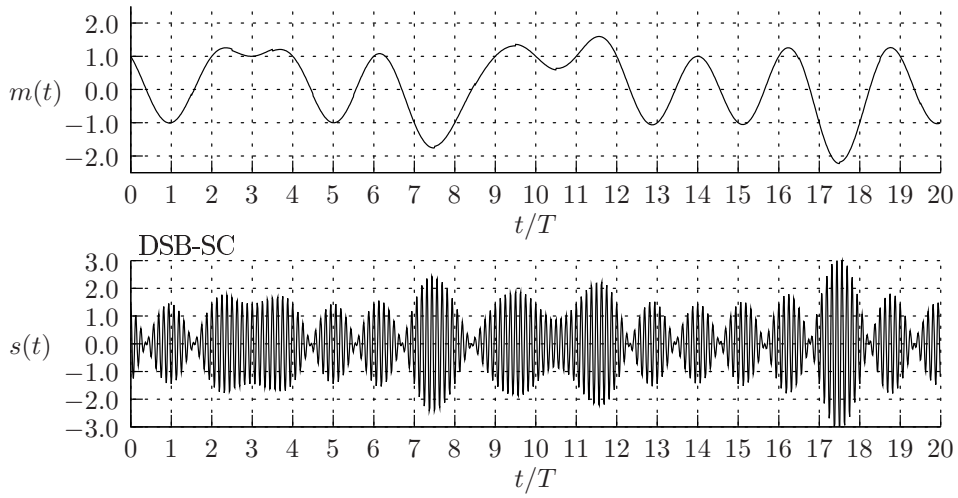


Figure 1.8: Top: message signal representing 20 information bits generated using raised-cosine pulses with $\beta = 0.1$. Bottom: DSB-SC signal, $s(t)$, normalized such that $< s^2(t) >= 1$.

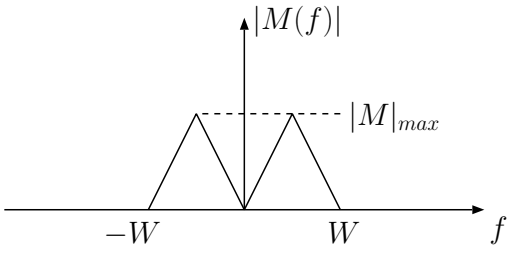


Figure 1.9: Two-sided spectrum of m(t).

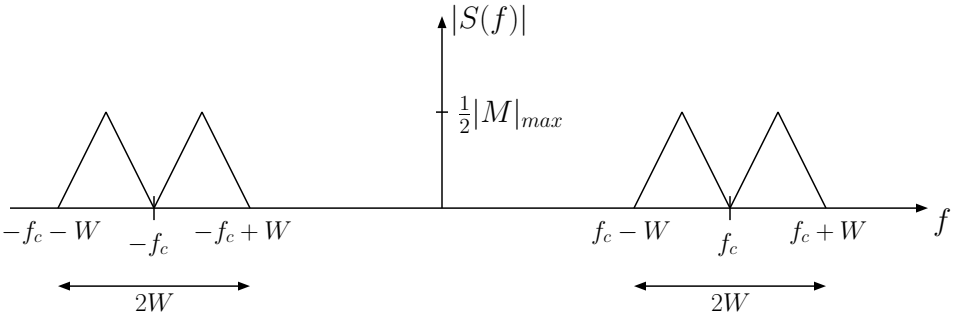


Figure 1.10: Two-sided spectrum of s(t).

overlap the carrier frequency is chosen so that $f_c > W$. It should be apparent that the DSB signal bandwidth is 2W (Hz) and therefore takes up twice as much of the spectrum as the original message signal.

Schematically DSB modulation can be represented as in Figure 1.11 where the modulated signal is shown being fed to an antenna for transmission.

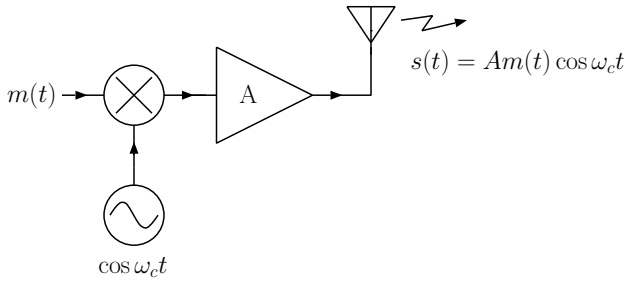


Figure 1.11: DSB modulation. The triangle represents a linear amplifier which scales the signal by the constant gain, A.

At the receiver, DSB can be demodulated using a process called *coherent demodulation* (or *synchronous demodulation*), which is illustrated in Figure 1.12. In this Figure, we have accounted for the fact that propagation between a transmitter and receiver will cause the carrier component of the signal to be phase-shifted with respect to the signal that was actually transmitted. We have ignored the fact that the received signal's amplitude will be reduced and that the message signal will be delayed.

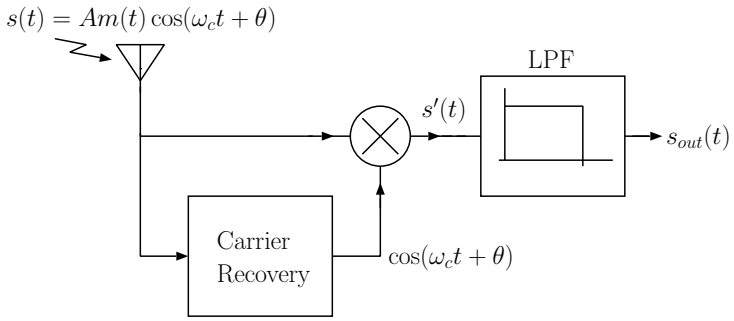


Figure 1.12: Coherent detection/demodulation.

The synchronous demodulation process requires the receiver to have a carrier recovery system which somehow recovers the phase-shifted carrier, $\cos(\omega_c t + \theta)$, from the DSB signal. Assuming that the carrier recovery system does its job perfectly, then it is easy to understand how the coherent demodulation process works. The signal after the multiplier, $s'(t)$, is the

product of $s(t)$ and the recovered carrier:

$$\begin{aligned} s'(t) &= s(t)\cos(\omega_c t + \theta) \\ &= Am(t)\cos^2(\omega_c t + \theta) \\ &= Am(t)\frac{1}{2}[1 + \cos(2\omega_c t + 2\theta)] \\ &= \frac{1}{2}Am(t) + \frac{1}{2}Am(t)\cos(2\omega_c t + 2\theta) \end{aligned} \tag{1.28}$$

The spectrum of $s'(t)$ is easily found using the modulation theorem and is sketched in Figure 1.13.

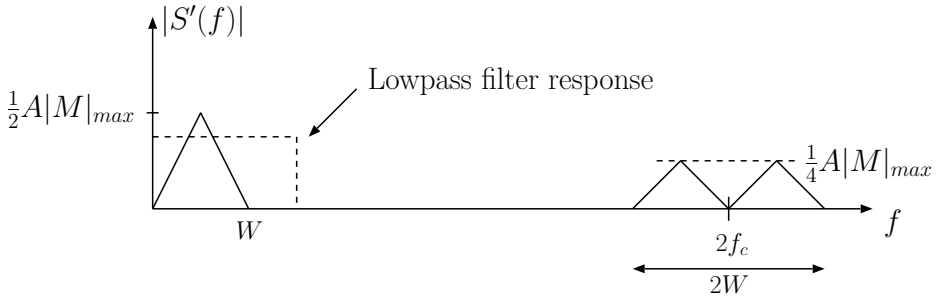


Figure 1.13: Spectrum of $s'(t)$. The ideal lowpass filter's cutoff frequency can be anywhere in the range $W < f_{cutoff} < 2f_c - W$.

The purpose of the lowpass filter is to reject the term $\frac{1}{2}Am\,(t)\cos(2\omega_c t + 2\theta)$ which corresponds to the part of the spectrum centered on $2f_c$. The output from the lowpass filter is

$$s_{out}(t) = \frac{1}{2}A\,m(t) \tag{1.29}$$

Note that for the purpose of this discussion we assumed that $f_c > W$, so that the spectra of the two terms in $s'(t)$ (after the multiplier) did not overlap. If the terms overlap, then it is not possible to separate the undesired component from the desired one with a lowpass filter and the demodulation scheme will not work. We also assumed that the receiver's carrier-recovery circuit was able to exactly recover the frequency and phase of the received carrier signal. Since these quantities must be estimated from a noisy received signal, we should consider the possibility that the carrier's frequency and phase might be imperfectly recovered, as in Figure 1.14. In this case

$$\begin{aligned} s'(t) &= Am(t)\cos(\omega_c t + \theta)\cos((\omega_c + \delta\omega)t + \theta + \delta\theta) \\ &= \frac{1}{2}Am(t)[\cos(\delta\omega t + \delta\theta) + \cos((2\omega_c + \delta\omega)t + 2\theta + \delta\theta)] \end{aligned} \tag{1.30}$$

After the lowpass filter

$$s_{out}(t) = \frac{1}{2}Am(t)\cos(\delta\omega t + \delta\theta) \tag{1.31}$$

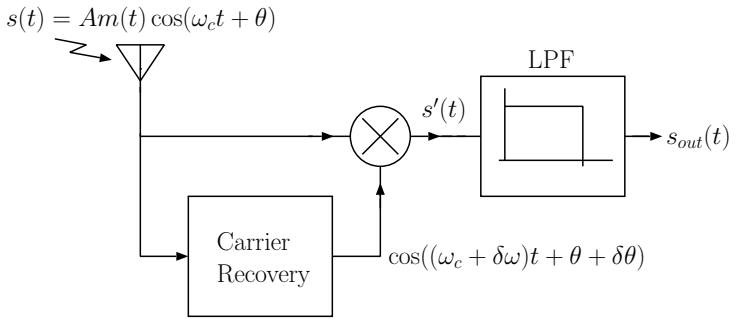


Figure 1.14: Recovered carrier slightly off-frequency and out-of-phase.

Comparing Equation 1.31 with Equation 1.29 we see that the effect of the frequency and phase error is to multiply the desired signal by $\cos(\delta\omega t + \delta\theta)$. This is unsatisfactory and would result in a distorted output. As an example, suppose that the message signal is a pure tone, i.e., $m(t) = \cos(\omega_m t)$. Then the output from a demodulator with frequency and phase error would be

$$\begin{aligned} s_{out}(t) &= \frac{1}{2}A\cos(\omega_m t)\cos(\delta\omega t + \delta\theta) \qquad (1.32) \\ &= \frac{1}{4}A[\cos((\omega_m + \delta\omega)t + \delta\theta) + \cos((\omega_m - \delta\omega)t - \delta\theta)] \end{aligned}$$

i.e., the output would consist of two tones separated in frequency by twice the frequency error. Clearly, a speech waveform that consists of the superposition of many "tones" would be seriously distorted unless the frequency error is extremely small. Thus it is important that $\delta\omega = 0$, i.e., it is necessary for the receiver's oscillator frequency to be exactly the same as that of the transmitter's carrier oscillator. Suppose $\delta\omega = 0$, then:

$$s_{out}(t) = \frac{1}{2}Am(t)\cos\delta\theta \qquad (1.33)$$

This is acceptable if $\delta\theta$ is constant and $\delta\theta \neq \pi/2$. It is not satisfactory if $\delta\theta$ varies with time or if $\delta\theta$ is close to $\pi/2$. The conclusion is that it is necessary for the receiver and transmitter oscillators to be frequency-synchronized and phase-locked to within a constant phase offset that is less than $\pi/2$. Carrier synchronization can be relatively easy to achieve if some synchronization information is sent by the transmitter.

An example of a familiar system that employs DSB-SC and provides separate carrier synchronization information is the FM stereo multiplex message signal where the difference between the left and right channel audio signals $(L(t) - R(t))$ DSB-SC modulates a 38 kHz subcarrier. This DSB-SC is summed with the monophonic signal (the sum of the left and right channels) along with a 19 kHz pilot tone derived by dividing the 38 kHz subcarrier frequency by two. The pilot tone is used by the receiver to reconstruct a 38 kHz carrier with the proper phase for demodulating the stereo information.

The analog color television signal also includes a component provided explicitly for the purpose of helping the receiver to achieve proper phase synchronization for coherent demodulation. In this case, coherent demodulation is necessary to extract the color information from the received video signal.

It is possible to design a carrier synchronization system that recovers the carrier from the received signal on its own without the aid of any extra synchronization information. It turns out that the missing carrier can be regenerated if the signal is first passed through a nonlinear device. For example, if the signal is passed through a square-law device, then the output will contain a component at twice the carrier frequency. This component can be extracted and its frequency divided by two to give the necessary carrier reference. This operation can be carried out using a phase-locked loop (PLL) in a circuit called a squaring loop or using a so-called Costas loop. (See Chapter 13.)

1.6.3 DSB with carrier

One way to provide the carrier synchronization information that is required to demodulate a DSB signal is to include an unmodulated carrier component in the transmitted signal. When a carrier component is included the the modulation is called DSB w/carrier. The difference between DSB-SC and DSB w/carrier is simply the addition of a carrier component to the spectrum in the latter case. The signal is represented mathematically by:

$$s(t) = (Am(t) + B)\cos(\omega_c t) \tag{1.34}$$

The signal can be written as the superposition of a DSB-SC signal and a carrier component:

$$s(t) = Am(t)\cos(\omega_c t) + B\cos(\omega_c t) \tag{1.35}$$

The spectrum of $s(t)$ is shown in Figure 1.15. The bandwidth occupied by DSB w/carrier

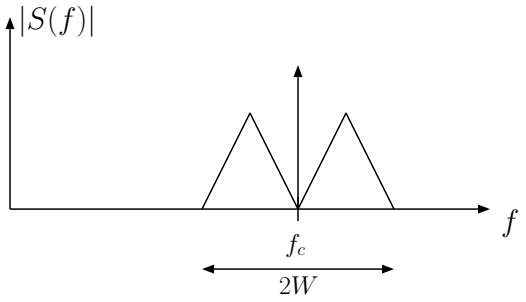


Figure 1.15: Spectrum of s(t) for DSB w/carrier.

is the same as DSB-SC and is 2W, where W is the highest frequency in the message signal, m(t). The coherent demodulator shown in Figure 1.12 can be used to demodulate DSB w/carrier. The presence of the carrier simplifies the implementation of the carrier recovery module. When a carrier component is present, the output of a coherent demodulator will contain a DC offset that is proportional to the strength of the received carrier component. The magnitude of the DC term can be used as a relative indication of received signal strength. It can also be used to control the gain of amplifiers preceeding the demodulator in order to keep the amplitude of the signal at the demodulator input relatively constant. Such a feedback system for controlling gain is called an *automatic gain control (AGC)* system.

1.6.3.1 DSB with full carrier

There is an important distinction between the cases where $Am(t)+B$ is allowed to become negative and where $Am(t)+B$ is constrained to always stay positive. If $Am(t)+B$ is always >0 the envelope of the modulated signal (defined by a line that connects positive peaks of $s(t)$) reproduces the message signal, $m(t)$. When $Am(t)+B>0$ for all t the signal is called *DSB with full carrier*. In this case it is possible to use a particularly simple demodulation circuit to recover $m(t)$ at the receiver. This circuit is known as an *envelope detector* and in its simplest form it consists of a half- or full-wave rectifier followed by a lowpass filter. Figure 1.16 (top) shows the same message signal that was given in Figure 1.8 and the bottom plot shows the corresponding DSB signal with full carrier. The signal in the bottom panel was produced from $m(t)$ in the upper plot by first forming the signal

$$s'(t) = (\frac{m(t)}{\max(|m(t)|)} + 1)\cos\omega_c t.$$

For this example the carrier frequency was chosen to be 8 times the signaling rate, i.e. $f_c = \frac{\omega_c}{2\pi} = \frac{8}{T}$. Dividing $m(t)$ by $\max(|m(t)|)$ ensures that the envelope $(\frac{m(t)}{\max(|m(t)|)} + 1) > 0$ for all t. The signal $s'(t)$ was then normalized to have a mean-square of 1.0. The normalized signal, $s(t)$, is shown in the bottom plot of Figure 1.16.

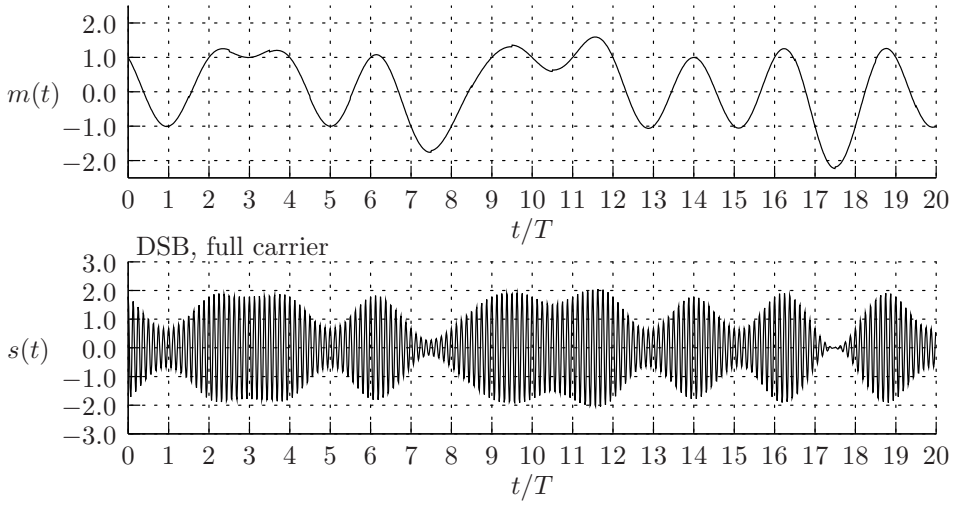


Figure 1.16: Top: message signal representing 20 information bits generated using raised-cosine pulses with $\beta = 0.1$. Bottom: DSB with full-carrier signal, $s(t)$, normalized such that $< s^2(t) >= 1$. Compare the bottom plot with the corresponding plot in Figure 1.8.

Figure 1.17 illustrates how the envelope detector demodulates a full-carrier DSB signal. The upper plot shows the result of passing $s(t)$ from Figure 1.16 through a full-wave rectifier, which produces $|s(t)|$. The lower plot shows the result of applying a low-pass filter to $|s(t)|$. Compare the lower plot of Figure 1.17 with the original message signal, $m(t)$, as shown in the upper plot of Figure 1.16 to verify that, aside from a constant offset and a scaling factor, the rectified and low-pass filtered signal is the same as $m(t)$.[2] It should be clear that

[2]The recovered signal is also slightly delayed in time due to the group-delay of the lowpass filter.

if the envelope detector is used in a situation where $Am(t) + B$ becomes < 0 (DSB with partial carrier), then the recovered envelope will only follow $m(t)$ during those times where $Am(t) + B > 0$. In such cases an envelope detector will recover a distorted version of $m(t)$.

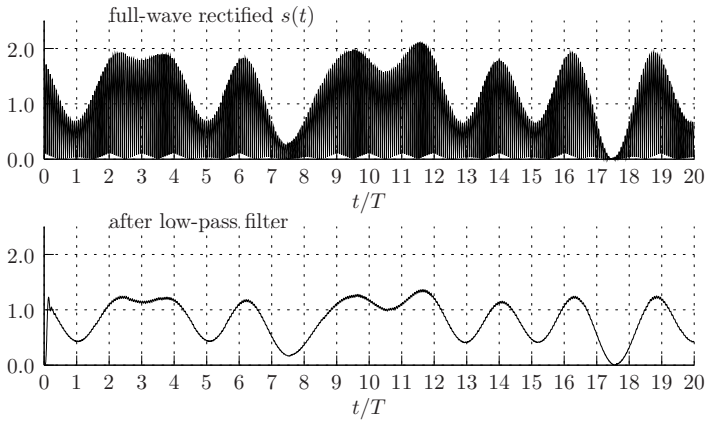


Figure 1.17: Upper plot shows $|s(t)|$, which is the result of passing $s(t)$ through a full-wave rectifier. The lower plot was obtained by passing the signal shown in the upper plot through a low-pass filter with cutoff frequency larger than the bandwidth of $m(t)$ (W) and smaller than $f_c - W$. The transient at the beginning of the lower plot is the start-up transient of the digital lowpass filter used to produce the simulated signal.

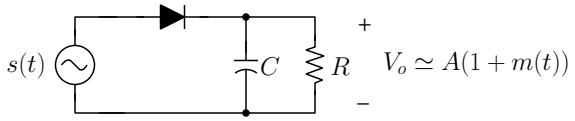


Figure 1.18: Practical envelope detector circuit.

The circuit of Figure 1.18 is commonly used to approximate the operation of an ideal rectifier/low-pass filter envelope detector. Assuming an ideal diode, the capacitor voltage will charge to the peak of the input signal $s(t)$ on a positive excursion of s(t), and then decay exponentially with time constant $\tau = RC$ thereafter until the input voltage rises and exceeds the capacitor voltage, at which time the capacitor voltage again follows the input signal to its peak. The resulting output voltage is an approximation to the signal envelope, i.e. $V_o \simeq Am(t) + B$. In order for the output to be a faithful reproduction of $m(t)$, it is necessary for the cut-off frequency of the lowpass RC filter to be larger than the bandwidth of $m(t)$ but smaller than the carrier frequency, i.e., $W \ll \frac{1}{2\pi RC} \ll f_c - W$.

It is also possible to use a square-law device instead of the rectifier, i.e., the rectifier is replaced with a square-law device as shown in Figure 1.19. Here the squaring operation is drawn to emphasize the similarity to a coherent detector as shown in Figure 1.12. Recall that the coherent demodulator can provide perfect recovery of the message signal. The square-law detector can be thought of as an approximation to the coherent detector, where

the input signal is used as the local carrier reference.

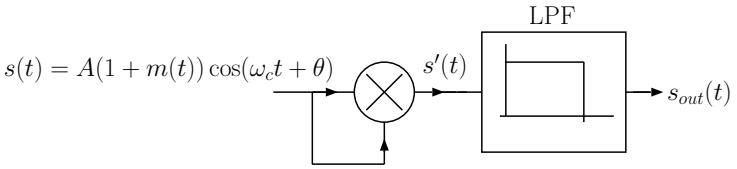


Figure 1.19: A square-law detector.

After the squarer the signal is

$$s'(t) = A^2(1 + 2m(t) + m^2(t)) \cos^2(\omega_c t + \theta) \tag{1.36}$$

$$= \frac{A^2}{2}(1 + 2m(t) + m^2(t)) \left[1 + \cos(2\omega_c t + 2\theta)\right]$$

The $\cos 2\,\omega_c t$ term is rejected by the lowpass filter. What remains is

$$s_{out}(t) = \frac{1}{2}A^2(1 \; + \; 2m(t) \; + \; m^2(t)) \tag{1.37}$$

The dc term can be removed with a coupling capacitor, leaving

$$A^2 \left[m(t) \; + \; \frac{1}{2}\, m^2\,(t)\right] \tag{1.38}$$

The first term is the desired output and the second term is an unwanted distortion term which will be small compared to the first term when $|m(t)| \ll 1$.

The circuit shown in Figure 1.18 can function as an envelope detector or square-law detector, depending on the magnitude of the input signal. If the peak input signal voltage is relatively small (i.e. less than approx. 26 mV peak), then the diode's exponential current-voltage characteristic can be expanded in Taylor series, in which case the leading nonlinear term will be a square-law term. For large input signals, with peak values exceeding the diode turn-on voltage, the circuit behaves more like an envelope detector based on a rectifier and low-pass filter.

1.6.4 Power efficiency for DSB signals

Let us assume that the DSB signal (with carrier) is applied across a 1Ω resistor and define the time-average power in the signal to be the average of the instantaneous power over a time interval that is long compared to all time scales of the message signal:

$$P_{\text{avg}} = \frac{1}{T}\int_{-T/2}^{T/2} s^2(t)dt.$$

For a DSB signal

$$P_{\text{avg}} = \frac{1}{T}\int_{-T/2}^{T/2} (Am(t) + B)^2 \cos^2(\omega_c t)dt.$$

Using brackets (<>) to denote the averaging operation, we have

$$\begin{aligned} P_{\text{avg}} &= < (Am(t)+B)^2 \cos^2(\omega_c t) > \\ &= < \tfrac{1}{2}(A^2 m^2(t) + 2ABm(t) + B^2)(1+\cos(2\omega_c t)) > \\ &\simeq \tfrac{1}{2}(A^2 < m^2(t) > + B^2) \end{aligned}$$

where we have used the fact that the long-time average value of the terms $m(t)$, $m^2(t)\cos(2\omega_c t)$ and $m(t)\cos(2\omega_c t)$ are negligibly small. The term $\frac{1}{2}A^2 < m^2(t) >$ is the contribution to average power from the message-signal modulation - this power is contained in the upper and lower sidebands of the DSB signal. The term $\frac{1}{2}B^2$ is the carrier's contribution to the average power. Define the modulation power efficiency of a DSB-with-carrier signal to be the fraction of the total transmitted power that resides in the information-bearing sidebands.

$$\frac{P_{\text{sidebands}}}{P_{\text{avg}}} = \frac{A^2 < m^2 >}{A^2 < m^2 > + B^2} \tag{1.39}$$

In the above equations $< m^2 >$ represents the mean square value of the modulating signal, $m(t)$. Let us assume that $m(t)$ varies symmetrically around zero. To ensure that $Am(t) + B > 0$ a full-carrier signal will have $B \geq A \max|m(t)|$. The largest power efficiency will occur when B is chosen to be as small as possible — so let us assume that $B = A\max|m(t)|$. The power efficiency in this case is

$$\frac{P_{\text{sidebands}}}{P_{\text{avg}}} \leq \frac{\frac{<m^2>}{\max|m(t)|^2}}{\frac{<m^2>}{\max|m(t)|^2}+1}$$

Since $\frac{<m^2>}{\max|m(t)|^2} \leq 1$, we conclude that the power efficiency of a full-carrier signal can be no larger than 50%. For example, suppose that $m(t)$ is a waveform consisting of contiguous rectangular pulses with amplitudes equal to ± 1. This corresponds to the digital modulation scheme known as "on-off" keying; when $m(t) = -1$, then $s(t) = 0$ and the transmitter is "off". In this case, $\frac{<m^2>}{\max|m(t)|^2} = 1$ and the modulation efficiency is 50%. Now suppose that $m(t)$ is something more like a typical analog message signal, i.e. suppose that $m(t)$ is a single sinusoidal tone, e.g., $m(t) = \cos(\omega_m t)$. In this case $\frac{<m^2>}{\max|m(t)|^2} = 0.5$ and the modulation efficiency is 33%. A realistic value for analog voice or music signals might be $\frac{<m^2>}{\max|m(t)|^2} < 0.25$, which results in modulation efficiency smaller than 20%.

In summary, in DSB with full carrier more than 50% of the average power is used to send the carrier which carries no message essential information — it is sent only to simplify the demodulation process. With realistic signals, as much as 80% of the transmitted power might be used to send the carrier. The remaining power is split between the two information-bearing sidebands. This seems like a significant waste of power. However, the advantage gained by using this scheme is that the demodulator in the receiver can be a very simple circuit consisting of as few as 3 passive components. If we attempt to make the scheme more efficient by reducing the carrier power such that the modulation index becomes greater than 1, then it is necessary to use coherent demodulation (as described for DSB-SC) to recover $m(t)$ and demodulator complexity rises dramatically.

1.6.5 Single-Sideband (SSB)

As we have seen in sections 1.6.1 and 1.6.3, DSB signals (with or without carrier) occupy a bandwidth $B = 2W$, twice the bandwidth of the message signal, $m(t)$. Since all of the information contained in $m(t)$ is in one sideband, the other sideband is redundant. This should be intuitively clear if we recall that the lower sideband in DSB results from translation of the negative-frequency part of the spectrum of $m(t)$ up into the positive frequency range. We also know that the negative-frequency part of the spectrum of $m(t)$ is related to the positive-frequency part through $M(-\omega) = M^*(\omega)$. Thus, all of the information is contained in one sideband, and DSB modulation schemes do not make the most efficient use of the radio frequency spectrum. It is more efficient to transmit only one sideband; either the upper sideband (USB) or lower sideband (LSB). Since only one sideband is included in the SSB signal, the bandwidth of a SSB signal is approximately W, which is the same as that of the message signal. Thus, SSB is spectrally efficient, since the bandwidth of the transmitted signal is the same as the bandwidth of the message signal.

Two methods for generating a SSB signal are described here.

The "filter" method:

The filter method is conceptually the simplest. First, the DSB signal is generated and then the unwanted sideband is filtered out, as in Figure 1.20.

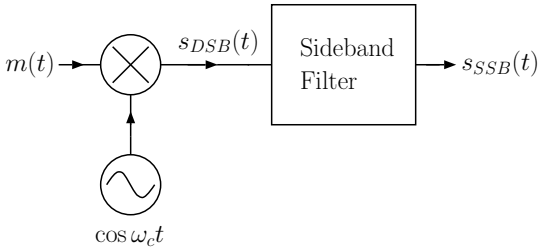


Figure 1.20: Filter method for generating an SSB signal.

This approach places severe constraints on the SSB filter, and it is necessary for $M(f)$ to have a low frequency gap so that the sideband filter can reject the unwanted sideband without significant attenuation of the desired sideband. The situation is illustrated in Figure 1.21. The two-sided spectrum of the message signal, $m(t)$, is shown on the left, and the one-sided spectrum of the DSB signal is shown on the right. The dotted line shows the frequency response of a realistic upper-sideband filter. The gap in the spectrum of $m(t)$ is required, because the sideband filter will have a transition region of finite width as shown in Figure 1.21. If the spectrum of $m(t)$ did not have a gap, i.e. if the spectrum extended down to DC with little or no attenuation, then the finite width of the sideband filter's transition region would result in unwanted attenuation of the desired sideband and/or leakage of the undesired sideband past the filter.

The primary advantages of the filtering method of SSB generation are simplicity and excellent rejection of the unwanted sideband. The disadvantages are the need for a low frequency gap in $M(f)$ and the need for highly selective unwanted sideband rejection filters. Filters that are suitable for rejecting the unwanted sideband are generally fairly expensive.

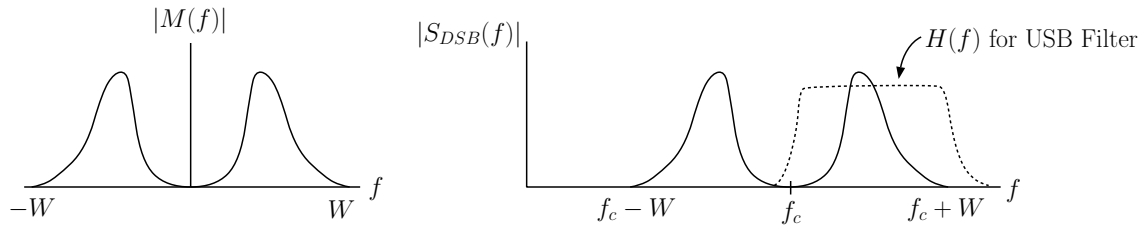


Figure 1.21: Left: two-sided amplitude spectrum of the message signal, $m(t)$. Right: One-sided amplitude spectrum of $s_{DSB}(t)$ (solid line) and frequency response of an upper-sideband filter (dotted line).

"Phasing" method:

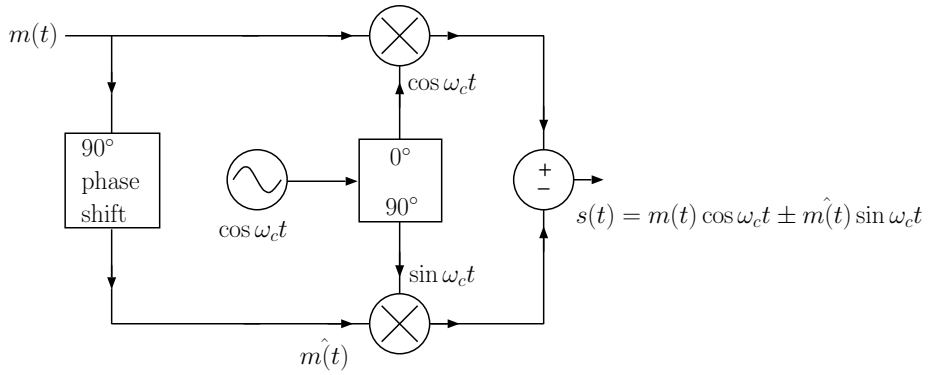


Figure 1.22: Phasing method for generation of an SSB signal.

A single sideband signal can be generated using a special type of filter (Hilbert transform filter) and a quadrature multiplexer/modulator. Given a message signal $m(t)$, define the Hilbert Transform of $m(t)$, denoted by $\hat{m}(t)$, to be the signal that is obtained by delaying each frequency component of $m(t)$ by 1/4 of a cycle, or $\pi/2$ radians. A filter that provides this delay function is called a Hilbert Transform filter. The Fourier transform of $\hat{m}(t)$ will have the same magnitude as the Fourier transform of $m(t)$, but the phase will differ by $-\pi/2$ (for positive frequencies) and $+\pi/2$ for negative frequencies. Thus, $\hat{m}(t)$ can be written in terms of $M(\omega)$ as follows:

$$\hat{m}(t) = \mathcal{F}^{-1}[-j \ \mathrm{sgn}(\omega) \ M(\omega)]$$

where $\mathrm{sgn}(\omega)$ is the signum function, defined by

$$\mathrm{sgn}(\omega) = \begin{cases} 1, & \omega > 0 \\ 0, & \omega = 0 \\ -1, & \omega < 0 \end{cases}$$

The single-sideband modulator forms the modulated carrier signal

$$s(t) = m(t)\cos\omega_c t \pm \hat{m}(t)\sin\omega_c t,$$

as illustrated in Figure 1.22. The upper sign corresponds to a lower-sideband (LSB) signal, and the bottom sign corresponds to an upper-sideband (USB) signal. To prove this, we can calculate the Fourier transform of $s(t)$. It is necessary to remember the following facts:

$$\mathcal{F}[\cos(\omega_c t)] = \pi[\delta(\omega - \omega_c) + \delta(\omega + \omega_c)]$$

$$\mathcal{F}[\sin(\omega_c t)] = j\pi[\delta(\omega + \omega_c) - \delta(\omega - \omega_c)]$$

$$\mathcal{F}[f(t)g(t)] = \frac{1}{2\pi} F(\omega) * G(\omega)$$

where $*$ denotes convolution. Now we can calculate the Fourier transform of $s(t)$:

$$\begin{aligned} S(\omega) &= \mathcal{F}[s(t)] \\ &= \tfrac{1}{2\pi}\left[M(\omega) * \pi[\delta(\omega - \omega_c] + \delta(\omega + \omega_c)] \mp j\mathrm{sgn}(\omega) M(\omega) * j\pi[\delta(\omega + \omega_c) - \delta(\omega - \omega_c)]\right] \\ &= \tfrac{1}{2}\left[M(\omega - \omega_c) + M(\omega + \omega_c) \pm [\mathrm{sgn}(\omega + \omega_c) M(\omega + \omega_c) - \mathrm{sgn}(\omega - \omega_c) M(\omega - \omega_c)]\right] \\ &= M(\omega - \omega_c)\ \tfrac{1}{2}[1 \mp \mathrm{sgn}(\omega - \omega_c)] + M(\omega + \omega_c)\ \tfrac{1}{2}[1 \pm \mathrm{sgn}(\omega + \omega_c)] \end{aligned}$$

Consider the upper sign for now, and examine the first term, which represents the positive-frequency part of $S(\omega)$. The term

$$\frac{1}{2}[1 - \mathrm{sgn}(\omega - \omega_c)] = \begin{cases} 1, & \omega < \omega_c \\ \frac{1}{2}, & \omega = \omega_c \\ 0, & \omega > \omega_c \end{cases}$$

is equal to zero above the carrier frequency, and hence removes the upper sideband of $M(\omega - \omega_c)$. If the lower sign is chosen, then the lower sideband will be multiplied by zero. Hence, when the upper sign is chosen, the part of the spectrum above the carrier frequency will be zeroed out, producing LSB. If the lower sign is chosen, the part of the spectrum below the carrier frequency will be zeroed out, producing USB. Exact cancellation of the unwanted sideband does not occur in practice, because of imperfections in the filters that produce the $\pi/2$ phase shift and because of amplitude imbalance between the two terms.

1.6.6 SSB Demodulation

Single sideband signals can be demodulated using coherent demodulation. The input signal is either a USB or LSB signal:

$$s(t) = A[m(t)\cos\omega_c t \pm \hat{m}(t)\sin\omega_c t] \tag{1.40}$$

With perfect carrier recovery, the output of the coherent demodulator is

$$s_{out}(t) = \mathrm{LPF}[s(t)\cos(\omega_c t)] = \frac{1}{2} A m(t).$$

If the recovered carrier has a phase error, i.e., if the recovered carrier is $\cos(\omega_c t + \delta\theta)$ then it is easy to show that

$$s_{out}(t) = \mathrm{LPF}[s(t)\cos(\omega_c t + \delta\theta)] = \frac{1}{2} A m(t)\cos\delta\theta \mp \frac{1}{2} A\hat{m}(t)\sin\delta\theta,$$

where $LPF[\,]$ represents the lowpass filter operation. The first term is the desired signal which has amplitude $\frac{1}{2}A\cos\delta\theta$, and the second term represents undesired "crosstalk" from the Hilbert transform signal $\hat{m}(t)$. If $m(t)$ is an analog voice signal, the crosstalk contribution from $\hat{m}(t)$ would not sound significantly different from the desired signal, $m(t)$. Intelligibility of voice does not depend critically on the phase relationship between the composite frequency components. For digital data transmission, however, the crosstalk term represents distortion and will degrade the performance of the system. It is necessary for the phase error $\delta\theta$ to be nearly zero in order for the desired signal to have an amplitude that is much larger than the crosstalk term. In a digital data transmission system some form of automatic carrier synchronization would be necessary.

When the message signal is an analog voice waveform, there is an important distinction between coherent demodulation as applied to demodulation of DSB and SSB. As we learned earlier, when demodulating DSB it is necessary to have frequency perfect synchronization in order to recover an intelligible signal. When demodulating an analog SSB voice signal, however, it turns out that some frequency error can be tolerated. The effect of a frequency error, $\delta\omega$, is to change the pitch of the received signal. For example, suppose that the modulating signal is a tone, i.e., $m(t) = \cos\omega_m t$. Then $s_{USB}(t) = A[\cos(\omega_m t)\cos(\omega_c t) - \sin(\omega_m t)\sin(\omega_c t)]$. Let's assume that at the receiver this signal is multiplied by $\cos[(\omega_c + \delta\omega)t]$. Then, after the multiplier we have the following:

$$\begin{aligned} s'(t) &= A\{\cos(\omega_m t)\cos(\omega_c t) - \sin(\omega_m t)\sin(\omega_c t)\}\cos[(\omega_c + \delta\omega)t] \qquad (1.41)\\ &= A\cos[(\omega_c + \omega_m)t]\cos[(\omega_c + \delta\omega)t]\\ &= \frac{1}{2}A\{\cos[(2\omega_c + \omega_m + \delta\omega)t] + \cos[(\omega_m - \delta\omega)t]\} \end{aligned}$$

We assume that both ω_m and $\delta\omega$ are $\ll \omega_c$ so that the lowpass filter will reject the first term. The output of the filter is then

$$s_{out}(t) = \frac{A}{2}\cos[(\omega_m - \delta\omega)t] \qquad (1.42)$$

Note that the effect of the frequency error is to cause the demodulated tone to be shifted in frequency by an amount equal to the frequency error. Contrast this to the DSB case where a frequency error resulted in *two* tones at the output for a single input tone. In SSB communication receivers used for analog voice reception the local oscillator can be a free-running oscillator (often called the beat frequency oscillator, or BFO). In practice the operator of the receiver would tune the BFO until the recovered speech signal is intelligible. The tuning is fairly critical, but frequency errors of several 10's of Hz can be tolerated. The audible effect is to shift the pitch of the recovered speech up or down by the frequency error, Δf.

Notice that coherent detection can be used to demodulate all of the forms of linear modulation that have been discussed so far - DSB-SC, DSB with carrier, and SSB. Only in the latter case is it permissible for the receiver oscillator to have a frequency error, and this applies only when the message signal is an analog voice signal. For DSB-SC and DSB with carrier it is necessary to have perfect frequency synchronization between the transmitter and receiver.

While SSB makes the most efficient use of the radio-frequency spectrum, this type of modulation places relatively severe demands on the power amplifier used to amplify the

signal for transmission, because the amplifier needs to be capable of delivering peak power that is significantly larger than the average power contained in the signal. This is illustrated in homework problem 10. In many cases, a better scheme for obtaining the same bandwidth efficiency as SSB with more reasonable peak to average power requirements is the quadrature multiplexing scheme discussed in the following section.

1.6.7 Quadrature Multiplexing

Quadrature multiplexing is not really a new modulation scheme, rather it is a method for modulating two baseband signals, each having bandwidth W, onto one carrier such that the total occupied bandwidth is $2W$, i.e. the bandwidth occupied by the modulated carrier is equal to the total bandwidth occupied by the original message signals. This doubles the bandwidth efficiency of single-channel DSB, wherein a baseband signal with bandwidth W is associated with an occupied bandwidth of $2W$. Of course, we already have encountered one modulation scheme (SSB) which produces a modulated carrier with the same bandwidth as the original message signal. In fact, using SSB, it is possible to transmit the upper sideband of one signal, $m_1(t)$, and the lower sideband of another signal, $m_2(t)$, thereby transmitting two signals using one carrier, and using the same total bandwidth as that occupied by the original baseband signals. This is sometimes done and the technique is called independent sideband or ISB modulation. Quadrature multiplexing offers a simpler way to accomplish the same thing while avoiding the complexities (filters and phasing networks) involved with generating the SSB signal, and that also results in a signal with more favorable peak to average power characteristics, The idea behind quadrature multiplexing it is to use two carriers with the same frequency but which differ in phase by 90 degrees. If the two message signals are $m_1(t)$ and $m_2(t)$, then the modulated carrier signal resulting from quadrature multiplexing is:

$$s(t) = m_1(t)\cos\omega_c t + m_2(t)\sin\omega_c t \tag{1.43}$$

Figure 1.23 is a schematic of a quadrature multiplexer/modulator that implements this scheme. The quadrature-multiplexed signal simply consists of two DSB-SC components,

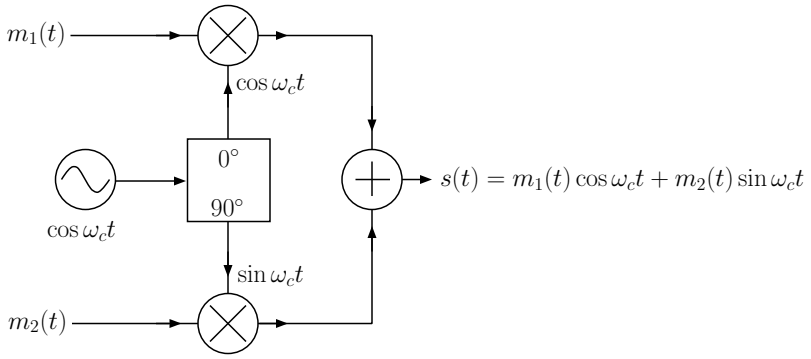


Figure 1.23: Quadrature multiplexer/modulator.

and the spectra of the two components will overlap, as in Figure 1.24. Because the spectra overlap it may look as if it would be impossible to recover the individual signals, $m_1(t)$ and $m_2(t)$, however it can be done if the signals are coherently demodulated as shown

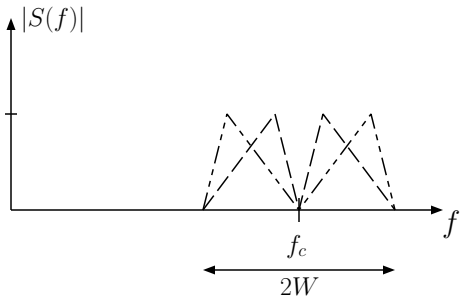


Figure 1.24: Two DSB-SC components with overlapping spectra.

in Figure 1.25, which shows a quadrature de-multiplexer. This system consists of two coherent demodulators, one (the upper branch) operates with an in-phase local oscillator, and produces an output, $I(t)$, that is called the in-phase signal component. The lower branch operates with a quadrature local oscillator, and produces an output, $Q(t)$, that is called the quadrature signal component. Provided that the local oscillator is properly synchronized with the carrier of the incoming signal, the in-phase output will be proportional to $m_1(t)$ and the quadrature output is proportional to $m_2(t)$. The demonstration that this demodulation scheme will work is left as an exercise.

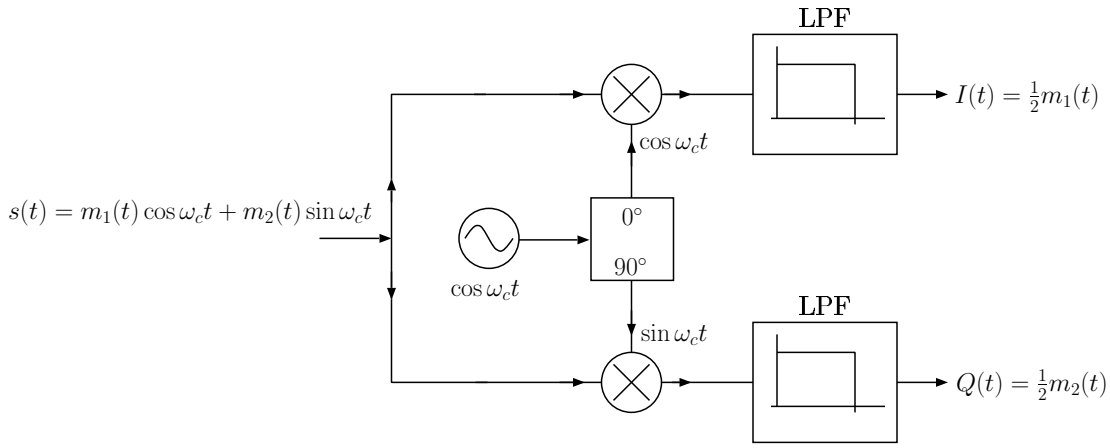


Figure 1.25: De-multiplexing of quadrature-multiplexed signals.

An application of quadrature multiplexing is the analog color television signal. The color, or "chrominance," information is contained in two signals, $I(t)$ and $Q(t)$, and is transmitted by quadrature multiplexing these signals onto a subcarrier with a frequency of approximately 3.58 MHz. In order to detect the color information at the receiver, it is necessary to have synchronization information for the receiver oscillator. This information is provided in the television signal by adding a short "burst" of the transmitter's 3.58 MHz signal into the

composite video signal at the end of each horizontal scan line. The signal is known as the "color burst." Although with this scheme the carrier information is present for only a small fraction of the time, it is still possible to use its information to lock an oscillator in the receiver to the transmitter's frequency and phase. This is usually done by employing a phase-locked loop which generates a continuous carrier and uses the periodic synchronization information to adjust the frequency and phase of the carrier.

1.6.8 Vestigial Sideband Modulation - VSB

Recall that SSB is the most bandwidth-efficient modulation, but it is difficult to get good low-frequency response in a SSB system because of the obstacles involved in the filtering or phasing methods of SSB generation. Low-frequency response is necessary in an analog television system so that video images with large areas of the same color and brightness can be transmitted. In a digital communication system, low frequency response is necessary in order to transmit pulses with non-zero mean. The hardware requirements are considerably relaxed if we allow a part ("vestige") of the unneeded sideband to be transmitted. If, in addition, a carrier component is transmitted, demodulation is simplified significantly. This vestigial sideband scheme is used for all commercial television transmissions, wherein the upper sideband and a vestige of the lower sideband are transmitted. Vestigial sideband is used for transmission of the video portion of analog television signals using the National Television Systems Committee (NTSC) system (see Figure 1.26a) and for transmittion of

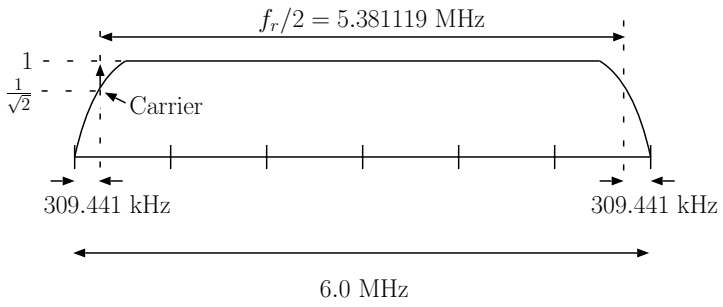


Figure 1.26: The spectrum of a television signal as transmitted for digital television (DTV) using the ATSC (Advanced Television Systems Committee) 8VSB system.

digital television (DTV) signals using the Advanced Television Systems Committee (ATSC) standard (Figure 1.26b). The ATSC modulation is called 8VSB.

ATSC message signals are constructed using pulses whose Fourier transform is the square-root of the raised-cosine pulses discussed earlier. In the receiver, the signal is filtered using a square-root raised-cosine response so that the received pulses will have a raised-cosine spectrum. A pulse can take on 8 possible amplitudes, so each pulse represents 3 bits of information. The fractional excess bandwidth of the pulses is $\beta = 0.0575$ and the signaling rate is $f_r = T^{-1} = 10.762238$ Mpulses/sec. Each pulse represents 3 bits of information so channel bit-rate is $3 \times 10.762 = 32.286$ Mbits/s. Redundant bits (coding) and overhead reduce the effective data rate to 19.3 Mbits/s. The signaling interval is

$T = 1/10.762 \times 10^6 \simeq 93\,\text{ns}$. The one-sided bandwidth of the pulses is $(1+\beta)f_r/2 = 5.6906$ MHz. A DSB signal generated from a message signal constructed from these pulses would have bandwidth $(1+\beta)f_r = 11.381$ MHz which exceeds the 6 MHz width of a single TV channel. As shown in Figure 1.26b, most of the lower sideband is removed to form the ATSC 8VSB signal spectrum. This is accomplished with a root raised-cosine rolloff that begins at approx. 309 kHz above the carrier frequency and falls to zero 309 kHz below the carrier frequency. Hence, the bandwidth occupied by the VSB signal is equal to the width of the upper sideband (5.6912 MHz) plus the width of the vestige of the lower sideband (309.4 kHz). The 8VSB signal is usually described as having an excess bandwidth of 11.5% (although the message signal pulses have 5.75% excess bandwidth) because the vestige of the lower sideband doubles the excess bandwidth.

1.7 Angle Modulation (Nonlinear Modulation)

We turn now the the other class of modulation schemes known as "angle modulation," or sometimes "nonlinear modulation." In this case the transmitted signal has the form

$$s(t) = A_c \cos(\omega_c t + \theta(t)) \tag{1.44}$$

where A_c is a constant. Thus the peak amplitude of the modulated signal is constant, but the phase angle is varied in response to the modulating signal. One important advantage of angle-modulation systems is their inherent insensitivity to amplitude fluctuations that may be present on the the received signal due to fading or noise. Since the angle-modulated signal has a constant amplitude envelope, the received signal can be passed through a hard limiter to remove amplitude fluctuations due to noise and fading; the modulation will not be distorted by this operation. The constant amplitude feature of these signals is also desirable because the peak to average power ratio is equal to one, allowing for efficient amplification of such signals to high power levels. Angle modulation is *nonlinear* in the sense that if $m_1(t)$ angle modulates a carrier to produce $s_1(t)$ (i.e. $m_1 \to s_1$) and $m_2 \to s_2$, then $m_1 + m_2$ does not result in the modulated signal $s_1 + s_2$. As a result, the spectrum of an angle-modulated signal is not related to the spectrum of the modulating signal in a straightforward way.

There are two basic types of angle modulation used for analog signals. These are distinguished by the way in which $m(t)$ is mapped onto the angle $\theta(t)$.

1. Phase Modulation (PM):
   The phase angle is a scaled version of $m(t)$, i.e.,

   $$\theta(t) = k_p m(t) \tag{1.45}$$

   where k_p is called the phase deviation constant.

2. Frequency Modulation (FM):
   The derivative of the phase angle is proportional to $m(t)$, i.e.,

   $$\frac{d\theta}{dt} = k_f m(t) \tag{1.46}$$

   where k_f is called the frequency deviation constant.

Note that in either case (PM or FM) the instantaneous frequency of the signal is given by

$$\omega_{inst} = \frac{d}{dt}(\omega_c t + \theta(t)) = \omega_c + \frac{d\theta}{dt} \tag{1.47}$$

In the case of FM we have

$$\omega_{inst} = \omega_c + k_f m(t) = \omega_c + \Delta\omega(t) \tag{1.48}$$

Thus, frequency modulation is generated simply by varying the frequency of the carrier signal in direct response to $m(t)$. For a general $m(t)$, then, the mathematical forms of phase- and frequency-modulated signals are as given in Equations 1.49 and 1.50:

$$\text{Phase} - \text{modulated}: \quad s(t) = A_c \cos(\omega_c t + k_p m(t)) \tag{1.49}$$

$$\text{Frequency} - \text{modulated}: \quad s(t) = A_c \cos(\omega_c t + k_f \int_{-\infty}^{t} m(t')dt') \tag{1.50}$$

1.7.1 Spectrum of Angle-modulated Signals

For nonlinear modulation there is no simple "modulation theorem" that provides a relationship between $M(f)$ and $S(f)$. We'll consider a simple case to gain some insight, and then we'll give a general rule of thumb for estimating the bandwidth of an angle-modulated signal.

1.7.1.1 Sinusoidal Modulation

Suppose that the message signal is a sine wave, i.e.,

$$m(t) = A_m \sin \omega_m t \tag{1.51}$$

Then

$$s_{PM}(t) = A_c \cos(\omega_c t + A_m k_p \sin \omega_m t) \tag{1.52}$$

$$s_{FM}(t) = A_c \cos(\omega_c t - \frac{A_m k_f}{\omega_m} \cos \omega_m t) \tag{1.53}$$

This example illustrates that except for a phase shift the PM and FM signals have the same functional form for the sinusoidal modulation case. For the purpose of deriving the spectrum of this type of signal it is sufficient to consider a function of the form

$$s(t) = A_c \cos(\omega_c t + \beta \sin \omega_m t) \tag{1.54}$$

where β is called the modulation index, which is simply the maximum phase deviation for either FM or PM. The modulation index can be related to the phase, or frequency, deviation constants through

$$\beta = A_m k_p \quad \text{(PM)} \tag{1.55}$$

$$\beta = \frac{A_m k_f}{\omega_m} \quad \text{(FM)} \tag{1.56}$$

Another interpretation for the modulation index comes from noting that the maximum instantaneous frequency deviation of the phase- or frequency-modulated signal is $\Delta\omega_{max} = \omega_m\beta$, so

$$\beta = \frac{\Delta\omega_{max}}{\omega_m} = \frac{\Delta f_{max}}{f_m} \tag{1.57}$$

Now the spectrum of $s(t)$ can be derived. The Fourier transform of $s(t)$ can be written as follows:

$$S(\omega) = \int_{-\infty}^{\infty} s(t)e^{-j\omega t}dt \tag{1.58}$$

$$= A_c\int_{-\infty}^{\infty} e^{-j\omega t}\cos(\omega_c t + \beta\sin\omega_m t)dt \tag{1.59}$$

This integral can be found in an integral table or table of Fourier transforms and the result is

$$S(\omega) = \pi A_c\sum_{n=-\infty}^{\infty} J_n(\beta)\delta(\omega-\omega_c-n\omega_m) + J_n(\beta)\delta(\omega+\omega_c+n\omega_m) \tag{1.60}$$

This result shows that the signal $s(t)$ has a spectrum which consists of a set of impulses that are located at the discrete frequencies $\omega = \pm(\omega_c \pm n\omega_m)$. The spectrum consists of an infinite number of sidebands that are separated from the carrier frequency by integer multiples of the frequency of the modulating tone, ω_m. The modulation index β is just a constant that describes the peak phase deviation of the signal. The function $J_n(x)$ is called the Bessel function of first kind of order n. These functions are tabulated, and for a given value of n, tables of $J_n(x)$ can be found in many mathematics reference books. The Bessel functions of negative order ($n < 0$) are related to Bessel functions of positive order through

$$J_{-n}(x) = (-1)^n J_n(x) \tag{1.61}$$

The terms $J_n(\beta)$ that appear in the expression for the spectrum can be thought of as coefficients that determine the strength of each of the impulses. Plots of the Bessel functions for $n = 0, 1, 2, 3, 4$, and 5 are shown in Figure 1.27.

Taking the inverse transform of the $S(\omega)$ gives an alternative form of the time-signal that may yield some insight. Note that each pair of delta functions at frequency $\pm\omega_n$ will yield a term of the form $2\cos\omega_n t$ when the inverse transform is applied. Then

$$s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)e^{j\omega t}d\omega = A_c\sum_{n=-\infty}^{\infty} J_n(\beta)\cos(\omega_c t + n\omega_m t) \tag{1.62}$$

Thus, the angle-modulated signal (with sinusoidal modulation) can be represented as an infinite superposition of cosinusoidal components having frequencies $|\omega_c + n\omega_m|$.

1.7.1.2 Example - sinusoidally modulated signal

Consider an angle-modulated signal of the form:

$$s(t) = A_c\cos(2\pi 20t + \sin(2\pi 2t)) \tag{1.63}$$

The carrier frequency for this signal is 20 Hz and the frequency of the modulating tone is 2 Hz. The peak phase deviation associated with the signal (β) is 1 radian, and the

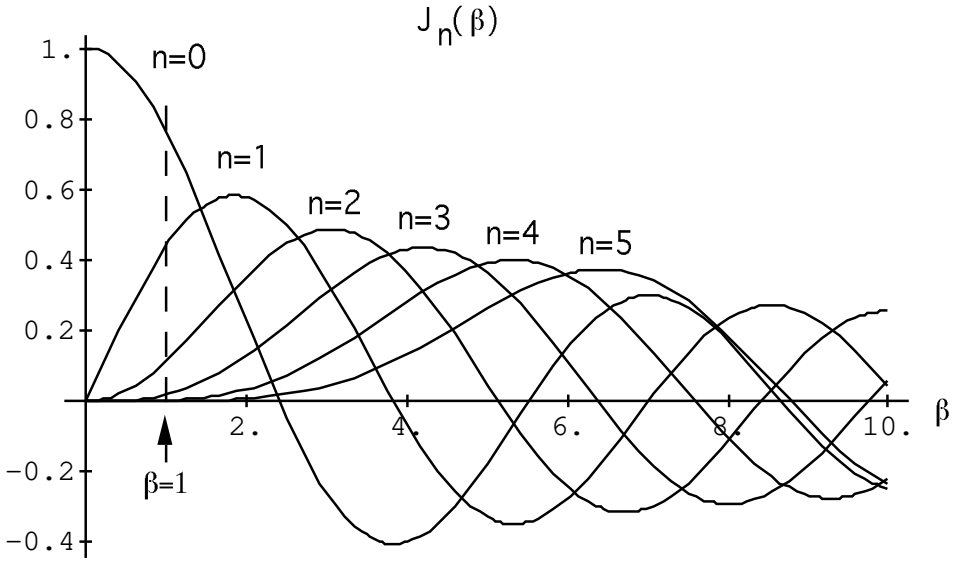


Figure 1.27: Bessel functions of order n for n=0,1,2,3,4,5.

peak frequency deviation is 2 Hz. According to Equation 1.62 this signal can be written as $s(t) = A_c \sum_{n=-\infty}^{\infty} J_n(1) \cos(2\pi(20 + 2n)t)$. The spectrum of this signal is easily determined by inspection of Equation 1.63. Each cosinusoidal component will contribute a pair of delta functions located at $f = \pm(20 + 2n)$ with amplitude given by the value of $J_n(1)$. The resulting line spectrum is shown in Figure 1.28. Note that for this example the amplitudes of the sidebands are smaller than that of the carrier. Figure 1.28 should be compared with Figure 1.27. The amplitude of each line in the spectrum can be determined by inspecting the values of $J_n(\beta)$ at $\beta = 1$ shown in Figure 1.27 by the dashed line. The value of $J_0(1)$ is the amplitude of the carrier component; the value of $J_1(1)$ gives the amplitude of the first sidebands, etc. If the peak phase deviation of an angle-modulated signal happens to correspond to a zero of one of the Bessel functions, then the corresponding component will not be present in the spectrum. A few of the zeros of the Bessel functions $J_0(x)$ and $J_1(x)$ are given in Table 1.5.

Table 1.5: Zeros of Bessel functions.

$J_0(x)$ has zeros at:	x= 2.4048	5.5201	8.6537	11.7915
$J_1(x)$ has zeros at:	x= 0.0000	3.8317	7.0156	10.1735

If, for example, the peak phase deviation is 5.5201 radians, the Fourier spectrum of the (sinusoidally modulated) angle-modulated signal will not have a carrier component. This phenomenon is useful for precise adjustment of the frequency deviation of FM transmitters. In practice the peak frequency deviation (Δf_{max}) allowed for an FM signal is fixed at some value which depends on the type of signal that is being generated. For example, a commercial FM broadcast signal is limited to a peak frequency deviation of 75 kHz. With sinusoidal modulation the peak phase deviation (β) of such a signal will depend on the frequency of

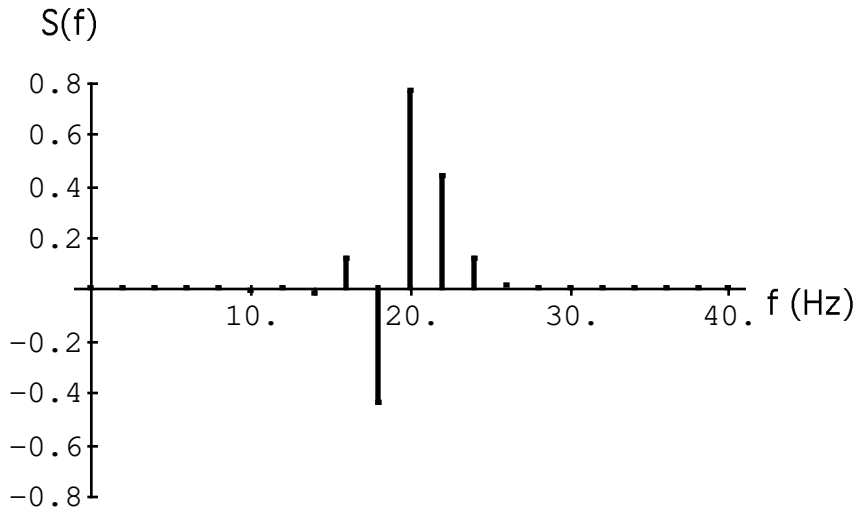


Figure 1.28: Line spectrum of angle-modulated signal with 20 Hz carrier and 2 Hz modulating tone.

the modulating tone through

$$\beta = \frac{\Delta f_{max}}{f_m} \tag{1.64}$$

Suppose an engineer wants to adjust the peak frequency deviation of a transmitter to some value, Δf_{max}. A precise adjustment can easily be made by using sinusoidal modulation and adjusting the frequency of the modulating tone to a value f_m such that, when Δf_{max} is set properly, the peak phase deviation β will take on a value corresponding to a zero of $J_0(x)$. The spectrum of the transmitted signal is then monitored on a spectrum analyzer, and the frequency deviation is adjusted until the carrier component vanishes. Similar adjustments could be made by using any of the sidebands instead of the carrier component.

1.7.2 Bandwidth of Angle-modulated Signals

Figures 1.29 through 1.32 show the magnitude of the Fourier Transform of a carrier that is angle modulated by a sinusoidal tone with $\beta = 1$, 2.4, 7, and 20. The spectrum actually consists of delta functions. The plots show the location and the relative amplitudes (absolute value) of the delta functions. Notice that in all cases the most significant sidebands are contained within the interval $f_c \pm (\beta + 1)f_m$. Outside of this interval the amplitude of the sidebands is relatively small and decreases rapidly with increasing separation from the carrier frequency. So a good approximation for the bandwidth of a carrier that is angle-modulated by a sinusoidal message signal is

$$BW \simeq 2(\beta + 1)f_m.$$

Since

$$\beta = \frac{\Delta f_{max}}{f_m},$$

the bandwidth can be written as

$$BW \simeq 2(\Delta f_{max} + f_m). \tag{1.65}$$

This is known as Carson's rule. It can be shown that the bandwidth given by equation 1.65 will contain at least 98% of the power associated with the angle modulated signal.

When the message signal is non-sinusoidal (as in almost all practical situations) and has bandwidth W, a reasonable approximation to the bandwidth results if Carson's rule is modified by replacing the tone frequency by the bandwidth of the message signal. Thus, for a non-sinusoidal message signal

$$BW \simeq 2(\Delta f_{max} + W).$$

When the modulating signal has finite bandwidth W can usually be taken to be the highest

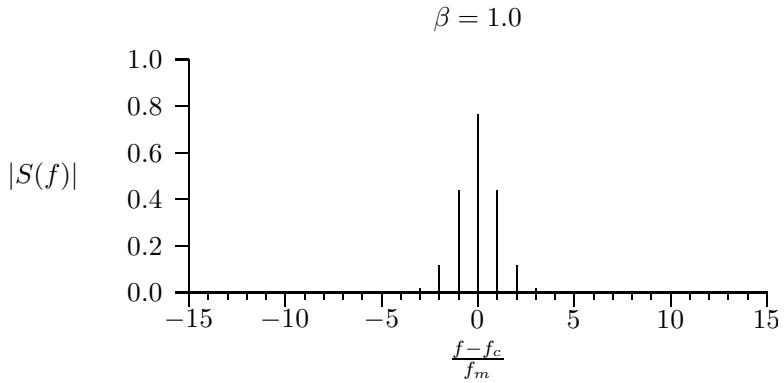


Figure 1.29: Spectrum of a carrier angle-modulated by a single tone with modulation index $\beta = 1$. Note that the frequency axis is the normalized frequency difference from the carrier frequency, i.e. $\frac{f-f_c}{f_m}$.

frequency that is contained in the signal. If the modulating signal has infinite bandwidth then W is chosen so that it contains most of the power in the signal.

Notice that when the peak frequency deviation is much smaller than the highest frequency in $m(t)$, i.e., if $\Delta f_{max} \ll W$, then Carson's Rule reduces to

$$BW \simeq 2W \tag{1.66}$$

When this approximation is valid, the modulation is called "narrowband" angle modulation, and the bandwidth is essentially the same as for AM or DSB. On the other hand, if $\Delta f_{max} \gg W$, the modulation is called "wideband" angle modulation and Carson's rule becomes

$$BW \simeq 2\Delta f_{max} \tag{1.67}$$

which states that the bandwidth is approximately twice the peak frequency deviation.

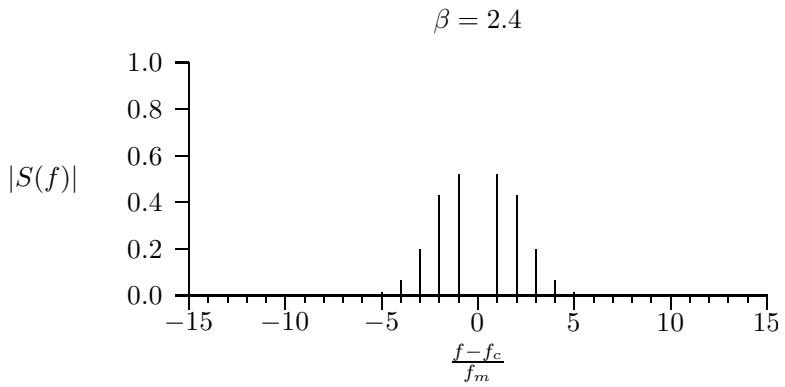


Figure 1.30: Spectrum of a carrier angle-modulated by a single tone with modulation index $\beta = 2.4$.

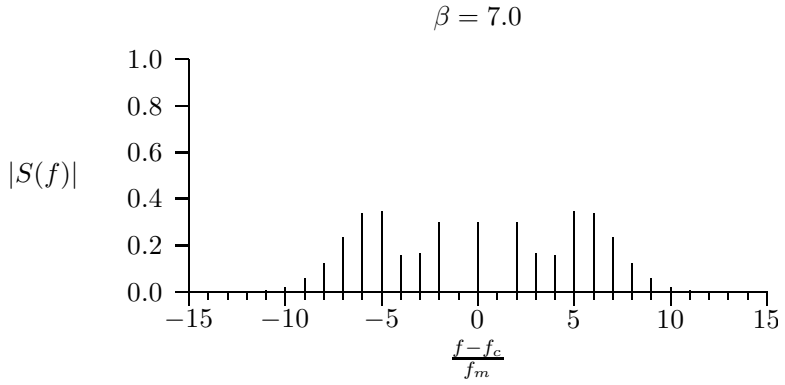


Figure 1.31: Spectrum of a carrier angle-modulated by a single tone with modulation index $\beta = 7$.

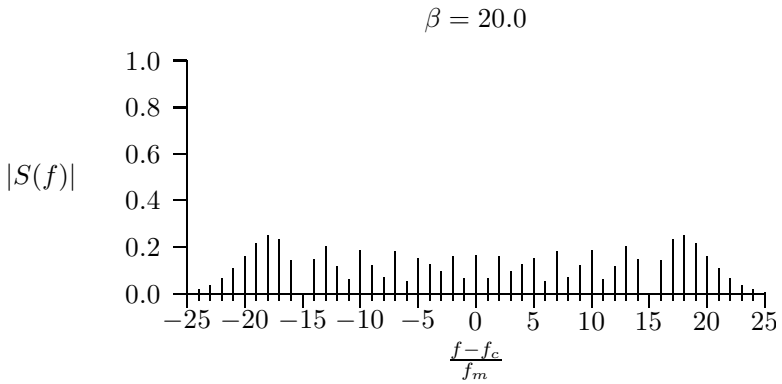


Figure 1.32: Spectrum of a carrier angle-modulated by a single tone with modulation index $\beta = 20$.

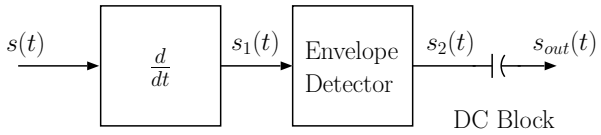


Figure 1.33: Idealized demodulator for FM signals.

1.7.3 Demodulation of FM

Most schemes for demodulating FM make use of the concept illustrated in the idealized demodulator shown in Figure 1.33. The input signal is of the form:

$$s(t) = A_c \cos(\omega_c t + k_f \int_{-\infty}^{t} m(t')dt') \tag{1.68}$$

After differentiation,

$$s_1(t) = \frac{d}{dt}s(t) = -A_c[\omega_c + k_f m(t)] \sin(\omega_c t + k_f \int_{-\infty}^{t} m(t')dt') \tag{1.69}$$

If $\omega_c + k_f m(t) \geq 0$, this looks very much like the AM signal. Recall:

$$S_{AM}(t) = A_c(1 + m(t)) \cos(\omega_c t + \theta) \tag{1.70}$$

After the envelope detector, approximately,

$$s_2(t) = A_c(\omega_c + k_f m(t)) \tag{1.71}$$

After dc blocking,

$$s_{out}(t) = A_c k_f m(t) \tag{1.72}$$

The differentiation operation can be approximated using a delay/difference scheme as shown in Figure 1.34. For the difference approximation to be a good approximation to the

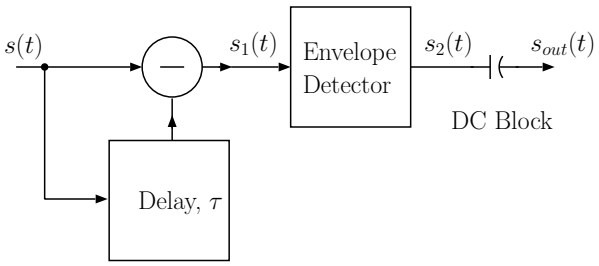


Figure 1.34: Approximation to the idealized FM demodulator.

derivative, the delay τ must be small compared to the period of the carrier oscillation, i.e. $\tau \ll 1/f_c$. This approach to demodulating FM yields a very small output signal, because the envelope of the signal after the differentiator is proportional to $\omega_c + k_f m(t)$. The second term is simply the instantaneous frequency deviation, so if the maximum deviation from the carrier frequency is $\pm f_{max}$, the ratio of peak to minimum envelope amplitude is

$$\frac{1 + \frac{\Delta f_{max}}{f_c}}{1 - \frac{\Delta f_{max}}{f_c}}.$$

Consider FM Broadcast, where $\Delta f_{max} = 75$ kHz and $f_c \sim 100$ MHz. The maximum to minimum envelope amplitude ratio is then 1.0015. Thus, the output from the envelope detector will be very small. For this reason, the differentiator/envelope detector is practical only for signals with relatively large deviations. When some distortion can be tolerated, much larger recovered signal can be obtained by replacing the differentiator with a single-resonator bandpass filter whose center frequency is deliberately offset from the carrier frequency. The idea is to place the carrier frequency on the skirts of the bandpass response, so that frequency deviations are translated into envelope variation. By employing a filter with fairly steep skirts (e.g. an LC filter with high Q), relatively large envelope variation can be obtained, even for small frequency deviations. This approach to demodulating FM is called *slope detection*.

An even better method for demodulating FM can be implemented by converting the instantaneous frequency deviation to relative phase deviations, and then using a phase-detector to convert the phase deviation to a voltage. A simple LC resonator can be used to convert the frequency deviation of the incoming signal into (relative) phase deviation. A circuit that accomplishes this is described in detail in Section 4.5.

1.8 Quadrature Modulation/Demodulation

In section 1.6.7 we pointed out that quadrature multiplexing can be used to send two message signals, $m_1(t)$ and $m_2(t)$, on one carrier, thereby doubling the bandwidth efficiency relative to single-channel DSB modulation. This property makes the quadrature multiplexer essentially a *universal modulator*, since it can be used to produce any type of modulation by properly choosing the two message signals. Consider a carrier signal that is both amplitude and angle modulated,

$$s(t) = A(t)\cos(\omega_c t + \theta(t)). \tag{1.73}$$

Straightforward application of the identity $\cos(a+b) = \cos a \cos b - \sin a \sin b$ allows equation 1.73 to be written as

$$s(t) = A(t)\cos\theta(t)\cos\omega_c t - A(t)\sin\theta(t)\sin\omega_c t$$

or

$$s(t) = m_1(t)\cos\omega_c t + m_2(t)\sin\omega_c t$$

with

$$m_1(t) = A(t)\cos\theta(t) \tag{1.74}$$

$$m_2(t) = -A(t)\sin\theta(t). \tag{1.75}$$

Therefore, to create a signal that is amplitude and/or angle modulated using a quadrature modulator, it is necessary to create the appropriate $m_1(t)$ and $m_2(t)$ signals according to 1.74 and 1.75, and apply these signals to a quadrature multiplexer, as shown in Figure 1.23.

Similarly, a quadrature de-multiplexer, combined with appropriate signal processing can serve as a *universal demodulator* whether or not a modulated signal was produced using a quadrature modulator. Refer to Figure 1.25, and note that when the input signal is an amplitude and/or angle modulated signal, the components $m_1(t)$ and $m_2(t)$ are given by equations 1.74 and 1.75. Thus, the output from the in-phase channel will be $I(t) = \frac{1}{2}m_1(t) = \frac{1}{2}A(t)\cos\theta(t)$, and the output from the quadrature channel will be $Q(t) = -\frac{1}{2}A(t)\sin\theta(t)$. The $I(t)$ and $Q(t)$ outputs can be provided as inputs to a signal processor, which extracts the envelope, $A(t)$, and/or the phase modulation, $\theta(t)$. The envelope is obtained by noting that

$$A(t) = 2\sqrt{I(t)^2 + Q(t)^2}. \tag{1.76}$$

The phase modulation is obtained from

$$\theta(t) = -\tan^{-1}(\frac{Q(t)}{I(t)}). \tag{1.77}$$

If the input signal is frequency modulated, then the demodulator must produce a signal that is proportional to $\frac{d\theta}{dt}$. Differentiating equation 1.77 gives the instantaneous frequency of the signal in terms of the I and Q outputs of the quadrature demultiplexer:

$$\Delta\omega(t) = \frac{d\theta}{dt} = \frac{-1}{I^2 + Q^2}(I\frac{dQ}{dt} - Q\frac{dI}{dt}). \tag{1.78}$$

If the signal envelope is constant, i.e. if the signal is purely angle modulated, then the term $I^2 + Q^2$ will be constant. This will be the case if the signal has been passed through a limiter before application to the quadrature demodulator. In that case, it is only necessary to compute the quantity $I\frac{dQ}{dt} - Q\frac{dI}{dt}$, which will be proportional to the instantaneous frequency deviation of the input signal, where the deviation is measured with respect to the local carrier reference.

In many applications, frequency demodulation is performed after the signal has been digitized. A discrete-time version of equation 1.78 can be written in terms of sampled-data sequences $\{I_n\}$ and $\{Q_n\}$ by replacing the time-derivatives with first-difference approximations, e.g. $\frac{dQ}{dt} \rightarrow \frac{1}{\Delta t}(Q_n - Q_{n-1})$. The result simplifies nicely and is easily implemented in a digital signal processor:

$$\Delta\omega_n = \frac{(I_n Q_{n-1} - Q_n I_{n-1})}{\Delta t(I_n^2 + Q_n^2)}. \tag{1.79}$$

It can be shown (see homework problem 14) that if the quadrature demodulator is to be used only for implementing envelope or instantaneous frequency recovery (as in equations 1.76 and 1.78), then it is not necessary for the local oscillator to be synchronized with the carrier of the incoming signal. In the case of envelope recovery, frequency or phase offsets have no effect, provided that the bandwidth of the LPFs is large enough to accomodate any frequency shifts in the I and Q branches. In the case of FM demodulation according to equation 1.78, frequency error in the local oscillator causes a DC offset in the output of the demodulator. The DC offset can be used as a tuning aid, or as an error signal in an *automatic frequency control* (AFC) loop, which would tune the local oscillator to minimize the frequency error.

1.8.1 Carrier Frequency and Phase Synchronization

The I and Q outputs from a quadrature demodulator can be used to form a control signal that corrects the instantaneous frequency of the local oscillator, for the purpose of synchronizing the local oscillator to that of the incoming signal's carrier. First, suppose that the local oscillator is perfectly synchronized with the incoming carrer, and that the input signal is a simple DSB signal so that $s(t) = m(t)\cos\omega_c t$. In this case, the I branch of the quadrature demodulator output contains the desired message signal, $I(t) = \frac{1}{2}m(t)$, and the Q branch output is zero. Notice that the product of the I and Q outputs is zero in this case. If the local oscillator is not synchronized to the incoming signal's carrier, then, in general, the local oscillator has both frequency and phase errors, i.e. the local oscillator is given by

$$\cos((\omega_c + \delta\omega)t + \delta\theta).$$

The quadrature local oscillator will then be

$$\sin((\omega_c + \delta\omega)t + \delta\theta).$$

If the input signal is DSB, then $s(t) = m(t)\cos(\omega_c t)$. The output from the I channel is:

$$I(t) = \frac{1}{2}m(t)\cos(\delta\omega t + \delta\theta).$$

The output from the Q channel is:

$$Q(t) = \frac{1}{2}m(t)\sin(\delta\omega t + \delta\theta).$$

The product of the I and Q channel outputs is

$$I(t)Q(t) = \frac{1}{8}m^2(t)\sin(2\delta\omega t + 2\delta\theta).$$

Notice that the IQ product is non-zero in the presence of frequency and/or phase errors. It turns out that if this IQ product is used to control the instantaneous frequency of the local oscillator, the resulting closed-loop feedback system will quickly drive the IQ product to a small value, and will reach an equilibrium where the frequency error $\delta\omega = 0$, and the phase error $\delta\theta$ is small. Such a feedback control system is called a Costas Loop, and is shown in Figure 1.35, where the IQ product is formed to generate a control voltage, $V_c(t)$. The local oscillator must be implemented as a Voltage Controlled Oscillator, so that the voltage V_c controls the instantaneous frequency of the oscillator.

In Figure 1.35, the feedback loop is opened by a switch in the control voltage path. It can be shown that, if the initial frequency error is small enough, when the switch is closed, the frequency error will be driven to zero, so that the control voltage becomes $V_c(t) = \frac{1}{8}m^2(t)\sin(2\delta\theta)$. The DC component of this control signal is $\overline{V_c} = \frac{1}{8}\overline{m^2(t)}\sin 2\delta\theta$. The loop adjusts itself so that this DC voltage is just large enough to tune the local VCO to the carrier frequency of the incoming signal. The adjustment is achieved by allowing a finite, static phase error, in the loop. The phase error will be just enough to provide the DC control voltage necessary to tune the VCO to the frequency of the incoming signal. If the VCO output frequency is a sensitive function of the control voltage, the required static phase offset will be very small, and the associated phase error will also be small. Practical implementations of the Costas Loop usually include a lowpass filter between V_c and the VCO to smooth the control voltage and control the dynamic response of the loop (the response to transient changes in the incoming signal's phase).

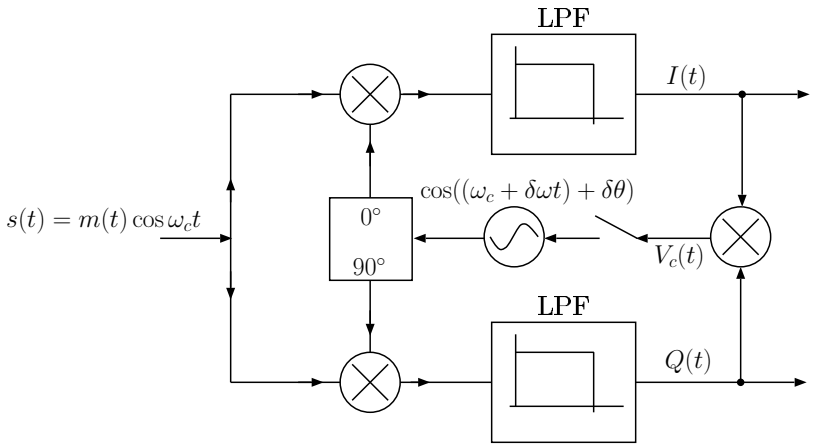


Figure 1.35: Costas Loop for carrier synchronization. The IQ product is used to control the instantaneous frequency of the local oscillator. When the switch is closed, the loop drives the IQ product to a small value.

1.9 References

1. Jordan, Edward C. and Keith G. Balmain, *Electromagnetic Waves and Radiating Systems*, Prentice Hall, 1968.

2. Proakis, J. and M. Salehi, *Communication Systems Engineering*, 2nd ed., Prentice-Hall, 2002.

3. Haykin, Simon, *Communication Systems*, 3rd ed., John Wiley and Sons, Inc., 1994.

4. Friis, Harald T., "A Note on a Simple Transmission Formula," *Proc. IRE*, pp. 254-256, May, 1946.

1.10 Homework Problems

1. One signal applied to the input of an ideal multiplier is

$$s_1(t) = \cos(1000\pi\, t) \;+\; \cos(1800\pi\, t) \tag{1.80}$$

The signal applied to the other input is

$$s_2(t) = \cos(7000\pi\, t) \;+\; \cos(20,000\pi\, t) \tag{1.81}$$

The output from the multiplier is applied to an ideal filter that passes only frequencies between 3.5 kHz and 10.0 kHz. List all frequencies that are present at the output of the filter.

2. An AM signal with full carrier can be represented as follows:

$$s(t) = A(1 + m(t))\cos(\omega_c t), \quad |m(t)| < 1 \text{ for all } t.$$

An envelope detector can be used to demodulate full-carrier AM. An ideal envelope detector is realized with a rectifier and a lowpass filter, as shown in Figure 1.36. The

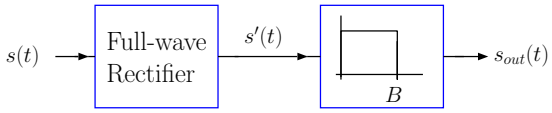


Figure 1.36: An ideal envelope detector, consisting of ideal full-wave rectifier and lowpass filter.

output of the full-wave rectifier can be written as $s'(t) = |s(t)| = s(t)p(t)$ where $p(t)$ is a square wave with the following properties:

- $p(t)$ is an even function,
  - $\max\{p(t)\} = 1$, $\min\{p(t)\} = -1$,
  - $p(t)$ has period $T = 2\pi/\omega_c$.

  (a) Express $p(t)$ in a Fourier series.

  (b) Sketch the amplitude spectrum of $s'(t)$. Assume that $m(t)$ has the amplitude spectrum shown in Figure 1.37. The bandwidth of $m(t)$ is assumed to be

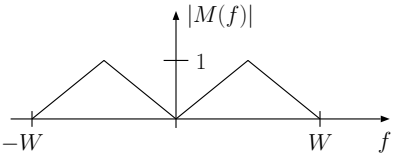


Figure 1.37: Amplitude spectrum of $m(t)$.

small compared to the carrier frequency, i.e. $W \ll f_c$.

(c) Specify the minimum and maximum possible cutoff frequency, B, for the low-pass filter that will allow $m(t)$ to be recovered at the output of the filter.

3. The message signal $m(t) = A\sin(\omega_m t)$ modulates a carrier signal given by $\cos(\omega_c t)$ where we assume $\omega_m \ll \omega_c$.

   (a) If the modulation is lower sideband (LSB), show that the modulated signal can be represented by

   $$f(t) = \frac{A}{2}\left[\sin(\omega_m t)\,\cos(\omega_c t) - \cos(\omega_m t)\,\sin(\omega_c t)\right] \tag{1.82}$$

   (Hint: Write down an expression for the DSB signal and manipulate it so that you have two terms – one for the upper sideband and one for the lower sideband. Then subtract the upper sideband component.)

   (b) Suppose that we attempt to demodulate this signal using a square-law detector. The output of a square-law detector is the square of the input signal. Assume that the detector is followed by a low-pass filter that removes all frequency components with frequencies larger than ω_c. Find the output of the square-law detector/low-pass filter. Can the modulating signal ($A\sin(\omega_m t)$) be recovered?

   (c) Consider the same problem as in 3b, except assume that the carrier is first reinserted, i.e., we add a term $K\cos(\omega_c t)$ to f(t) to give:

   $$f'(t) = \frac{A}{2}\left[\sin(\omega_m t)\,\cos(\omega_c t) - \cos(\omega_m t)\,\sin(\omega_c t)\right] + K\cos(\omega_c t) \tag{1.83}$$

   Assume that this signal is applied to the square-law detector/low-pass filter. Can the modulating signal be recovered? How will the output of the demodulator depend on the amplitude of the reinjected carrier? Note: In practice the oscillator that provides the reinjected carrier for SSB demodulation is called the beat-frequency oscillator (BFO).

4. Find the instantaneous frequency as a function of time for the following signal:

   $$s(t) = 10\cos[2\pi(10^8 t - 10^4 t^2) - 4\pi]$$

5. Consider an angle-modulated signal of the form

   $$s(t) = A\cos(\omega_c t + 5\cos\omega_m t + 15\cos 3\omega_m t) \tag{1.84}$$

   where $\omega_c = (2\pi)\,10^7\ s^{-1}$, and $\omega_m = (4\pi)\,10^3\ s^{-1}$.

   (a) What is the peak frequency deviation of this angle-modulated signal? Give your result in kHz. Be careful, you need to find the absolute maximum value of a function with several sub-maxima.

   (b) Estimate the bandwidth of this signal using Carson's rule. You will need to determine the bandwidth of the baseband signal (W) for this calculation. Choose the smallest bandwidth that contains all of the power in the baseband signal. Give your result in kHz.

(c) Suppose that the given s(t) is the result of frequency modulation by a baseband signal denoted by m(t). Specify the function m(t). (Note: You may assume that the frequency deviation constant, k_f, is equal to 1.0.)

6. Consider a signal m(t) with Fourier transform M(ω) where

$$M(\omega) = \int_{-\infty}^{\infty} m(t)e^{-j\omega t}dt \quad \text{and} \quad m(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} M(\omega)e^{j\omega t}dw \tag{1.85}$$

(a) Consider a particular M(ω) which is sketched in Figure 1.38. You may assume that M(ω) is a real function for this problem. Sketch the Fourier transform

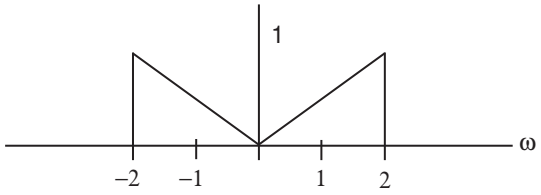


Figure 1.38: M(ω)

(spectrum) of the DSB signals

i. $m(t)\cos(t)$
ii. $m(t)\cos(6t)$

(b) Suppose we attempt to recover m(t) from the DSB signal using coherent demodulation. Can m(t) be recovered in both cases? Why or why not?

7. An angle-modulated signal can be represented by

$$s(t) = A\cos[\omega_c t + \theta(t)] \tag{1.86}$$

Consider the instantaneous phase deviation $\theta(t)$ of an angle-modulated signal as shown in Figure 1.39.

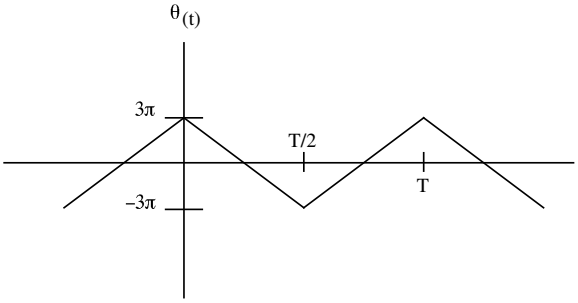


Figure 1.39:

(a) Suppose that this signal is the result of frequency-modulating the carrier signal. Sketch the message signal m(t). You may assume that $k_f = 1$.

(b) Suppose that T = 1 ms. What is the peak value of the instantaneous frequency deviation of this signal?

(c) Estimate the bandwidth of the angle-modulated signal using Carson's Rule. Explain any assumptions or approximations that you make.

8. FCC regulations state that the deviation of an FM-broadcast transmitter must be no greater than ± 75 kHz. Consider an FM transmitter with carrier frequency $f_c = 100$ MHz and unknown peak frequency deviation. To measure the actual deviation of the transmitter, the engineer modulates the carrier with a single tone ($m(t) = \cos\omega_m t$) such that the (unknown) peak deviation stays constant, i.e. the peak deviation does not depend on the tone frequency.

   The frequency of the tone is constrained to be in the range 20 Hz to 20 kHz to stay within the audio bandwidth of the modulator. The frequency of the tone is adjusted while the spectrum of the modulated signal is monitored. When the tone frequency is decreased from 20 kHz, it is found that the carrier component of the spectrum vanishes at $f_m = 10.40$ kHz, the first sidebands vanish at $f_m = 6.52$ kHz, the carrier vanishes at $f_m = 4.53$ kHz, the first sidebands vanish at f_m= 3.56 kHz, etc.

   (a) What is the actual peak deviation of the transmitter? Express your result in kHz.

   (b) Next, the engineer sets the frequency of the modulating tone to a fixed value, f'_m, and increases the frequency deviation of the transmitter. As the deviation is increased from the original value found in part a, the amplitude of the carrier component of the spectrum is observed to go to zero twice. When the amplitude of the carrier component reaches zero for the second time, the engineer knows that the deviation has been set to ± 75 kHz. What is f'_m in kHz? (Hint: you may assume that 12.5 kHz $< f'_m <$ 20 kHz.)

   (c) If the engineer had stopped at the first zero, what would the deviation be? (Hint: See Table 1.5 for a list of the first few zeros of $J_0(x)$ and $J_1(x)$.)

9. A tone ($\cos 2\pi f_m t$) frequency modulates a carrier. The tone frequency, f_m, is varied while the frequency deviation, Δf_{max} is held constant. Consider the Fourier amplitude spectrum of the modulated signal. Suppose the carrier component of the spectrum (i.e., the delta function at the carrier frequency) is observed to vanish when the tone frequency, f_m, is equal to 2.350 kHz and at 3.684 kHz.

   (a) What is the peak frequency deviation of the signal? Express your result in kHz.

   (b) Give two more tone frequencies that would make the carrier component vanish. The zeros of the first two Bessel functions are given below:
   $J_0(x)$ has zeros at x = 2.4048, 5.5201, 8.6537, 11.7915
   $J_1(x)$ has zeros at x = 0.000, 3.8317, 7.0156, 10.1735

10. When designing or specifying a transmitter it is necessary to consider the "peak envelope power" (PEP) as well as the average power of signals that are to be delivered to

the antenna for transmission. The "envelope power" of a modulated carrier signal s(t) is a time-varying quantity defined to be the average of the instantaneous power over one cycle of the RF carrier, i.e. the envelope power (assuming 1 Ω impedance) is:

$$P_{\text{env}}(t) = \frac{1}{\tau_c}\int_{t-\tau_c/2}^{t+\tau_c/2} s(t')^2 dt' \ , \ \tau_c = \frac{2\pi}{\omega_c}$$

It is usually a good approximation to assume that the envelope of the signal is effectively constant over the duration of a carrier cycle, in which case the envelope function can be pulled out of the integral. The peak envelope power (PEP) is the peak value of $P_{\text{env}}(t)$.

The average power is defined as the average of the instantaneous power over a time interval that is long compared to all time scales of the message signal:

$$P_{\text{avg}} = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} s(t')^2 dt'$$

Consider the message signal given below:

$$m(t) = \frac{4}{\pi}\sum_{n=1,3,5\ldots}^{N}\frac{\sin(n\omega_m t)}{n}$$

This series is the truncated Fourier series for a square wave.

(a) Write down the series expression for the Hilbert transform, $\widehat{m}(t)$. The Hilbert transform of $m(t)$ ($\widehat{m}(t)$) is obtained by delaying every frequency component of $m(t)$ by 90°. Note, if $m(t) = \sin(\omega t)$ then $\hat{m}(t) = -\cos(\omega t)$.

(b) Plot the waveforms $m(t)$ and $\hat{m}(t)$ for N=1,5,9,13,17.

(c) If the band-limited square wave is transmitted using SSB, the envelope (amplitude) of the resulting modulated carrier signal will be $A(t) = \sqrt{m(t)^2 + \hat{m}(t)^2}$. The envelope power of the modulated carrier signal will be proportional to $A^2(t)$. For each N, plot $A^2(t)$ and estimate the peak to average ratio, PAR $\equiv \frac{\text{PEP}}{P_{avg}}$, numerically.

- Note: This exercise illustrates why we would not consider using single-sideband (SSB) for a digital communication system that uses pulse amplitude modulation with rectangular pulses. You should find that $\hat{m}(t)$ exhibits large values wherever $m(t)$ has a large derivative. The peak values of $\hat{m}(t)$ increase as the derivatives of $m(t)$ increase (with increasing N, in this problem). The large peaks in $\hat{m}(t)$ cause large PEP values, requiring the transmitter to be able to deliver large instantaneous output powers. Notice that the peak-to-average power ratio is 1.0 for a DSB signal, $m(t)\cos\omega_c t$, when the message signal is constructed from "square" pulses. Thus, by going to SSB we reduce the occupied bandwidth by a factor of two - but the price is increased peak-to-average power ratio. Note that the bandlimited "pulses" (small N) do not have very sharp edges and they produce a smaller PAR. Similarly, a shaped pulse with rounded edges (such as a raised-cosine pulse) does not exhibit the large amplitude excursions, and hence the peak-to-average power ratio is not very large.

11. Estimate the bandwidth of these signals:

    (a) $s(t) = 10\cos[2\pi 10^8 t - 2\ \cos(2\pi 10^4 t + \pi/4)]$

    (b) $s(t) = 50\cos[2\pi 10^8 t]\ \cos[2\pi 10^4 t + \pi/4]$

12. Consider the following angle modulated signal:

$$s(t) = 50\cos(\omega_c t + 5\sin(2\pi 1000 t))$$

    (a) Determine the peak phase deviation in radians.

    (b) Determine the instantaneous frequency deviation in Hz.

    (c) Determine the peak frequency deviation in Hz.

    (d) Use Carson's rule to calculate the bandwidth of $s(t)$. Express your result in kHz.

    (e) Suppose $s(t)$ is passed through a bandpass filter that passes all frequencies within 3.5kHz of the carrier frequency f_c without attenuation and rejects all other frequency components (i.e. the bandpass filter has a rectangular transfer function that is centered on f_c with a bandwidth of 7 kHz). Express the power at the output of the filter as a percentage of the power in $s(t)$. You may want to use Mathematica, Matlab, or similar program to obtain the numerical values of $J_n(\beta)$ that you will need for this part.

13. A tone $m(t) = \cos(2\pi f_m t)$ frequency modulates a carrier. The tone frequency, f_m, is initially set to 400 Hz and is slowly increased. As the tone frequency is increased the frequency spectrum of the modulated signal is monitored. The sidebands at $f_c \pm f_m$ are observed to vanish when f_m is equal to 491.5 Hz and again at 712.7 Hz. What is the peak frequency deviation of the signal?

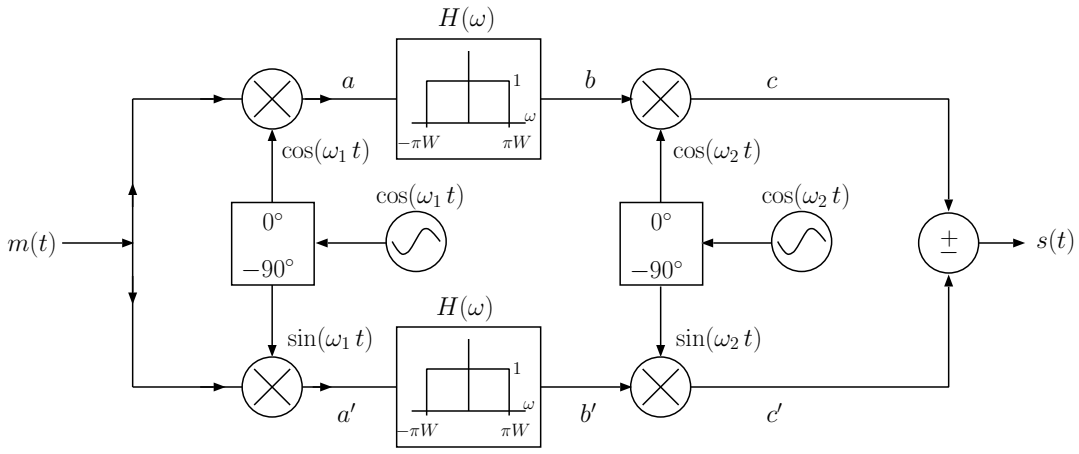
14. The signal at the input to a quadrature demodulator is

$$s(t) = A\cos(\omega_c t + k_f \int_{-\infty}^{t} m(t')dt').$$

    The quadrature demodulator produces two output signals, $I(t)$ and $Q(t)$, by multiplying $s(t)$ with in-phase and quadrature local oscillator signals, respectively. Each multiplier is followed by a low-pass filter that rejects double-frequency terms. A signal processor takes the outputs of the quadrature demodulator and produces the output signal $s_{out}(t) = I(t)\frac{dQ(t)}{dt} - Q(t)\frac{dI(t)}{dt}$. In general, the local oscillator signals employed within the quadrature demodulator have frequency and phase error, i.e. the in-phase and quadrature local oscillator signals are $\cos((\omega_c + \delta\omega)t + \delta\theta)$ and $\sin((\omega_c + \delta\omega)t + \delta\theta)$, respectively.

    (a) If $\delta\omega = 0$ and $\delta\theta = 0$, show how $s_{out}(t)$ depends on the message signal $m(t)$; i.e., derive an expression for $s_{out}(t)$ that is in terms of $m(t)$.

    (b) Now allow the frequency and phase errors to be finite. How do frequency and phase errors affect $s_{out}(t)$?



15. Two methods for generating SSB (and VSB) signals were introduced in this chapter — the filter method and the phasing method. There is a third method, called the "Weaver method" which is illustrated in the Figure. The bandwidth of the message signal is assumed to be W Hz ($2\pi W$ s^{-1}). In the frequency domain, the signal at point a can be written as

$$S_a(\omega) = \frac{1}{2}[M(\omega - \omega_1) + M(\omega + \omega_1)].$$

At point b, the signal becomes

$$S_b(\omega) = S_a(\omega)H(\omega) = \frac{1}{2}[M(\omega - \omega_1) + M(\omega + \omega_1)]H(\omega).$$

And at point c

$$\begin{aligned} S_c(\omega) &= \tfrac{1}{2\pi} S_b(\omega) * \pi[\delta(\omega - \omega_2) + \delta(\omega + \omega_2)] \\ &= \tfrac{1}{4}[M(\omega - \omega_2 - \omega_1) + M(\omega - \omega_2 + \omega_1)]H(\omega - \omega_2) + \\ &\quad \tfrac{1}{4}[M(\omega + \omega_2 - \omega_1) + M(\omega + \omega_2 + \omega_1)]H(\omega + \omega_2). \end{aligned}$$

(a) Find the corresponding expressions for the frequency-domain signals at points a', b', and c'.

(b) If the upper (+) sign is chosen for the combiner, determine the frequency-domain signal at the output of the combiner. Simplify as much as possible - some terms should cancel. If $\omega_2 \gg \omega_1$, specify ω_1 in terms of W if the output is to be a SSB signal. Determine whether the modulator yields an upper sideband (USB) or lower sideband (LSB) signal. Hint: the term $H(\omega - \omega_2)$ is 1 for $\omega_2 - \pi W \leq \omega \leq \omega_2 + \pi W$ and is zero elsewhere.

(c) Now assume that the lower (-) sign is chosen. Determine ω_1 such that the modulator yields a SSB signal. Is the output USB or LSB?

(d) If the message signal has bandwidth $W = 5.0$ MHz, and you want to generate a USB signal with carrier frequency at 1250 MHz and upper cutoff frequency at

1255 MHz, specify the correct sign (+ or -) to use in the combiner and the values of $f_1 = \omega_1/(2\pi)$ and $f_2 = \omega_2/(2\pi)$. Give your results in MHz.

(e) For the same message signal as in part (d), specify the sign to use in the combiner and f_1 and f_2 if it is desired to generate LSB with the carrier frequency at 1250 MHz and the lower cutoff frequency at 1245 MHz.

16. DSB-SC is used to transmit a message signal, $m(t) = \sum a_n p(t - nT)$, where $T = 1\,\mu s$ and $p(t)$ is a raised-cosine pulse with $\beta = 0.5$. Calculate the bandwidth occupied by the transmitted signal.

17. Consider the following signal:

$$s(t) = 10\sin[2\pi(10^9)t + 2\cos(3\pi 10^5 t)].$$

(a) Find the instantaneous frequency (in Hz) as a function of time for this signal.

(b) What is the peak instantaneous frequency deviation of $s(t)$? Express your result in kHz.

(c) Estimate the bandwidth of $s(t)$. Express your result in kHz.

(d) The fractional bandwidth of $s(t)$ is defined to be the signal bandwidth divided by the carrier frequency. What is the fractional bandwidth of $s(t)$?

(e) Suppose that $s(t)$ is the result of frequency modulation by a message signal denoted by $m(t)$. Specify the function $m(t)$. You may assume that the frequency deviation constant, k_f, is equal to 1.0.

(f) In general, the Fourier spectrum of $s(t)$ will contain delta functions with components at the carrier frequency, f_c, and at frequencies separated from the carrier frequency by multiples of some frequency f_m. What is f_m? Give your answer in kHz.

(g) Refer to Figure 1.27 to answer this question. In the Fourier amplitude spectrum of $s(t)$, ($|S(f)|$), denote the strength of the delta function at f_c by a_0, and the strength of the delta functions at $f_c \pm nf_m$ by a_n. Sort the list $\{a_0, a_1, a_2, a_3, a_4\}$ in order of decreasing amplitude.

Chapter 2

Receivers

2.1 Introduction and Historical Progression

The history of radio communication begins at the end of the 19'th century, when Italian inventor Guglielmo Marconi developed the first practical radio communication system using spark-gap radio transmitters and a nonlinear circuit element, called a "coherer", as a detector at the receiving end of the link. His most famous accomplishment was demonstration of transatlantic wireless communications, in 1902, when the Morse-code letter "S" (dot dot dot) was transmitted from England, and received in Newfoundland. Wireless telegraphy transmission using spark-gap transmitters used the broadband damped oscillations generated when the DC current in an LC circuit is suddenly interrupted. Spark-gap telegraphy was the primary wireless communications technology in use throughout World War I. Meanwhile, beginning in 1906, the technology required for the next phase of radio communication development was put into place when Lee De Forest patented the "audion", a device created by inserting a third electrode into the vacuum diode, creating a "triode" vacuum tube. This 3-terminal device could be used to amplify signals, and could be made to oscillate, thus serving as a very stable source of continuous-wave (CW) high-frequency signals. By 1915, the American Telegraph and Telephone Company (AT&T) had developed a high-power amplitude-modulated (AM) radio transmitter based on the triode, and in that year, for the first time in history, voice transmissions from Arlingon, VA were heard in Paris, France, and in Honolulu, HI. In 1920, the first commercial AM radio stations began operating. For much more information on the fascinating early history of radio communication, the book *The Science of Radio*, by Paul J. Nahin, is highly recommended.

Our purpose in this chapter is, first, to give a brief outline of the historical development of radio receivers, beginning with the simple, passive, tuned detector, and second, to present a fairly detailed discussion of the modern superheterodyne receiver.

2.1.1 Tuned Detector/Demodulator

The purpose of a receiver is to select, amplify, and demodulate a modulated-carrier signal, while ignoring the many other signals which will be picked up by the antenna, but are not of interest. The earliest receivers employed a single tuned circuit as a filter in front of an envelope detector implemented with a "crystal detector", a point-contact type of diode. Voice transmissions were amplitude modulated with full-carrier, so an envelope detector was

sufficient to demodulate the signal. The block diagram of such a *tuned-detector* receiver is shown in Figure 2.1. The tuned-detector is a passive circuit, and the audio power available

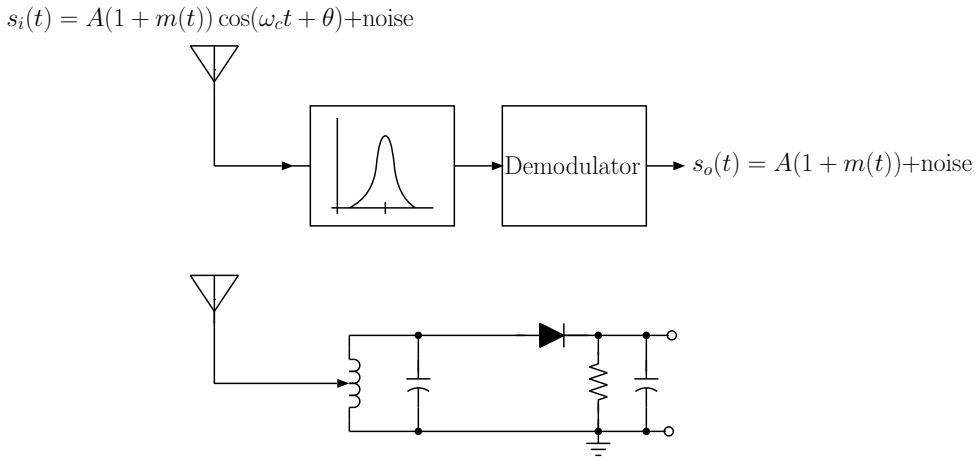


Figure 2.1: Top: block diagram of a tuned-detector receiver. Bottom: typical implementation. This type of receiver is also known as a "crystal set", because early implementations of the diode junction were implemented with the point-contact junction between a wire and crystal such as galena (lead sulfide) or carborundum (silicon carbide).

from the detector output must be provided by the energy collected by the antenna. Large antennas are necessary to collect enough signal, even from relatively close transmitters, to provide an output that is audible even with sensitive headphones. Selection of the desired signal, and rejection of unwanted signals, is the responsibility of the resonant circuit, which acts as a filter, and which must be tuned to pass the desired carrier frequency of interest. It is easy to improve the ability of a tuned-detector to receive weak signals by adding an amplifier to the system, but this does not solve the other fundamental problem with this receiver, which is the fundamental lower limit to the bandwidth that can be achieved with simple filters based on inductor/capacitor resonators. This limitation will be discussed in more detail in section 2.2. For the time being, it suffices to say that an LC filter, whose center frequency must be adjustable, is limited to a bandwidth no smaller than approximately 1% of its center frequency. While tunable LC filters with smaller bandwidths can be designed, the attenuation (loss) of such filters soars as the design bandwidth is decreased below about 1%. In the early days of radio, carrier frequencies were low, so fractional bandwidths of signals were relatively high, and the 1% lower limit on bandwidth was not a serious problem. As shorter wavelengths came into use, the fractional bandwidth occupied by signals became smaller than 1%, and some means of obtaining narrower bandwidth was needed.

2.1.2 Tuned Radio Frequency (TRF) Receiver

One very popular receiver architecture that was employed in the 1920's for AM broadcast receivers was sort of a brute-force approach to obtaining high gain and narrow bandwidth. The *tuned-radio-frequency* (TRF) receiver utilized a cascade of several tuned amplifiers,

each having a bandpass response, as shown in Figure 2.2, where each tuned amplifier is represented by separate filter and amplifier blocks. If the filter/amplifier stages are identical, and if the amplifiers provide an effective buffer between the filters, then the overall transfer function of the cascade is equal to the transfer function of a single filter/amplifier stage raised to the n'th power, where n is the number of stages. This architecture provides high gain and narrower bandwidth than can be obtained with a single filter. The TRF receivers were fairly difficult to tune properly, as each of the tuned circuits had to be tuned to the carrier frequency of interest. If only one of the filters was mis-tuned, the overall gain of the receiver would drop to a very low value. Another problem with the TRF receiver was the tendency for the system to oscillate. This problem is a manifestation of the "too much gain in one box" syndrome, which occurs whenever reverse isolation (the attenuation between the output and the input) of an amplifier chain does not exceed the forward gain by a comfortable margin. Leakage from the output back to the input provides a feedback path, which can lead to oscillation. In early receivers, the internal feedback within the triode vacuum tubes used as amplifiers provided enough reverse coupling to cause oscillation unless the forward gain was kept small. A technique called "neutralization" was developed to cancel the internal feedback of the active devices, and the TRF receiver employing neutralization became known as the neutrodyne receiver. The triode-based neutrodyne was followed by a second generation TRF receiver based on the tetrode vacuum tube, which had almost no internal feedback and did not require neutralization. The tetrode-based TRF receiver was available through the early 1930s.

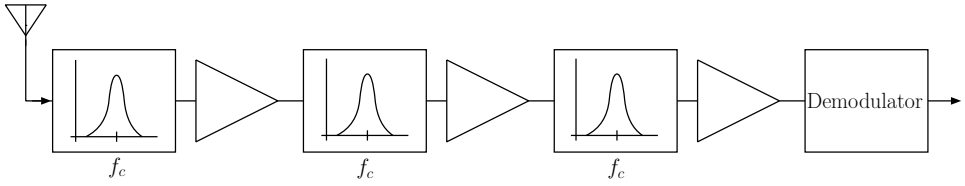


Figure 2.2: Tuned-radio-frequency (TRF) receiver architecture.

2.1.3 Regenerative Receiver

The TRF receiver became a practical commercial device only when the cost, and availability, of triode vacuum tubes made these devices available to the general public. Even before the dawn of AM broadcasting, and the subsequent popularity of the TRF receiver, a particularly ingenious concept was developed by Edwin Howard Armstrong, in 1912, which allowed him to develop a receiver with high gain, and narrow bandwidth, using only a *single* triode. Armstrong, an undergraduate at Columbia University at the time, used positive feedback to increase the gain, and narrow the bandwidth of a single-stage tuned amplifier. A block diagram of the system is shown in Figure 2.3. The transfer function of the system between the antenna terminals and the demodulator input is

$$\frac{s_o(\omega)}{s_i(\omega)} = \frac{GF(\omega)}{1 - AGF(\omega)}. \tag{2.1}$$

Without the positive feedback, the transfer function would be equal to the numerator of equation 2.1. The gain of the filter/amplifier would be $GF(\omega_o)$, and the fractional bandwidth

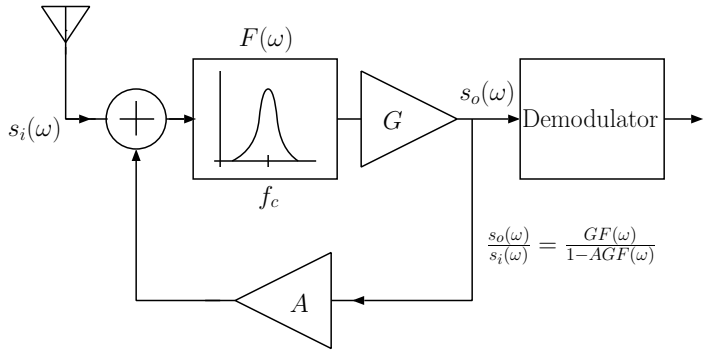


Figure 2.3: A regenerative receiver employing a regenerative amplifier in front of a demodulator.

of the system would equal to that of the filter. With feedback, the gain at the center frequency is $GF(\omega_o)/(1 - AGF(\omega_o))$, which can be made arbitrarily large if $AGF(\omega_o)$ is allowed to approach 1 (from below!). The positive feedback that is built into the regenerative amplifier is precisely what is necessary to build an oscillator. Note that if the quantity $AGF(\omega_o)$ is allowed to become equal to one, the transfer function becomes infinite, which may be interpreted as a condition where finite output signal occurs with zero input, in other words, the system oscillates and becomes a radio-frequency signal source, rather than an amplifier. Armstrong's regenerative receiver operates the feedback loop just below the threshold of oscillation, where the denominator of equation 2.1 is very small. Hence, the gain of the system is very large. It turns out that if the filter is a simple LC resonator, then the gain-bandwidth product of the system is independent of the value of $AGF(\omega_o)$. Hence, when the gain is very large, the bandwidth of the system can be very small. This provides a means to obtain extremely high gain and extremely small fractional bandwidths, using only a single active device.

The drawback to the regenerative receiver is the fact that the desired operating point is where $AGF(\omega_o)$ is close to one. If something happens to cause $AGF(\omega_o)$ to become equal to, or greater than, one, then the system becomes unstable, and oscillates. When the system is oscillating, it no longer acts as an extremely high gain and narrow bandwidth amplifier for small signals delivered by the antenna. In the oscillating mode, the regenerative receiver essentially becomes a self-oscillating mixer, which does not provide any narrow bandwidth filtering action. The difficulty associated with adjusting the filter and feedback to achieve high gain and narrow bandwidth made this receiver relatively difficult to use. It is also inherently very sensitive to changes in component values and circuit layout, so that it is rarely used in modern systems. Nevertheless, in the 1920s, there was an intense competition between Westinghouse, who had acquired the rights to Armstrong's regenerative patent, and a group of radio manufacturers that owned the rights to the neutrodyne patent.

An interesting, and even more exotic, variant of the regenerative receiver concept can be found in the superregenerative receiver. Keeping in mind that the regenerative receiver is most sensitive (in the sense that it has the highest gain) when it is operated at the threshold between stability and instability (oscillation), the superregenerative receiver is designed so that the quantity $AGF(\omega_o)$ is actually larger than 1. Thus, oscillation starts and builds

up immediately after power is applied to the system. Superregenerative receivers have a built-in circuit that senses when oscillation starts (this is easily detected by changes in the bias point of the active device) and, when oscillation is detected, the gain of the system is momentarily reduced, so that the oscillation is quenched. After the oscillation is quenched, the gain is restored, and oscillation builds up again. This cycle is repeated with a period that is short compared to the shortest time scale associated with the modulation of the signal that is being received. Thus, the system operates at extremely high gain for short periods of time (while oscillation is building up) during each cycle, effectively providing samples of the desired signal at a sampling rate that is adjusted to be high enough to allow the detector to recover the modulation. Superregenerative receivers can still be found today in devices such as garage door openers and radio controlled toys, where the primary design requirement is absolute minimum cost.

2.1.4 Genesis of the Superheterodyne Receiver

Although the regenerative and TRF receivers dominated the consumer radio market into the early 1930s, their replacement, and the architecture that is dominant today, was patented by Edwin Armstrong in 1917, and is called the superheterodyne receiver. After his stint at Columbia, Armstrong became a member of the U.S. Army Signal Corps during World War I. Involved with efforts to find a way to detect enemy airplanes from a distance, he knew that it might be possible to detect the electromagnetic emissions from the spark plugs in the engine. The problem was that the emissions were strongest at the (then) unusually high frequencies above a few MHz. Triodes of the day had very little gain at such high frequencies, so Armstrong hit on the idea of employing the heterodyne principle to shift the high-frequency signals to a lower frequency, where they could be more efficiently amplified and filtered. Once he decided to incorporate the heterodyne concept into his receiver, it became possible to heterodyne any signal of interest, regardless of its frequency, to a fixed *intermediate frequency* (IF). Selective filtering, and high-gain amplification could be done at the fixed IF.

In contrast to the TRF and regenerative receivers, the new heterodyne-based receiver did not require a tunable, high-gain, narrow-bandwidth, filter/amplifier. Instead, the high gain, narrow bandwidth filter/amplifier only had to operate at the fixed IF which was a relatively low frequency. This made it much easier to optimize the IF filters and amplifiers for high gain and narrow bandwidth. A further advantage was obtained because converting the desired radio frequency (RF) signal to a lower IF does not change the absolute bandwidth of the signal. For example, a signal that occupies a bandwidth of 10 kHz at a carrier frequency of 10 MHz occupies a fractional bandwidth of 0.01/10=0.001, or 0.1%. When that same signal is heterodyned to an IF of, say 400 kHz, it retains its 10 kHz bandwidth, so its fractional bandwidth becomes 10/400=0.025, or 2.5%, a comfortably large fractional bandwidth. Receivers based on Armstrong's superheterodyne concept were originally produced by Radio Corporation of America (RCA) who, as a result, dominated the commercial radio market by 1930. Before proceeding with a detailed overview of the superheterodyne receiver, we shall first discuss the properties of practical filters.

2.2 Characteristics of Practical Filters

Losses in the components used to implement filter networks set a lower limit on the bandwidth that can be achieved in a bandpass filter. For a bandpass filter with center frequency

f_o, a common measure of the bandwidth of the filter is the frequency spacing between the -3dB points on the filter transfer function, denoted by Δf_{-3dB} in Figure 2.4. The dimen-

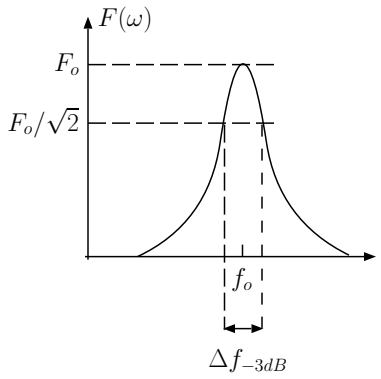


Figure 2.4: Definition of -3dB bandwidth.

sionless fractional bandwidth, $\Delta f_{-3dB}/f_o$, of a filter implemented with a single LC resonant circuit is $1/Q$, where Q is the so-called loaded Q of the resonant circuit. In many cases, the resonant circuit Q is limited by the Q of the inductor. Capacitor losses are usually small compared to the losses in inductors. For miniature inductors, Q's may be limited to values in the range 10-100, whereas inductors implemented using spirals of metallization in integrated circuits typically have very low Q, often $<$10. The fractional bandwidth of a filter implemented using an LC resonator with Q of 100 is 0.01, which means that the minimum bandwidth of an LC filter is about 1% of the center frequency. If physically large inductors can be tolerated, it is possible to achieve Q's as large as several hundred, even approaching 1000, and fractional bandwidth's significantly smaller than 0.01. Such high-Q inductors are rarely practical for use in modern systems where small size, light weight, and low cost are primary considerations. Generally speaking, however, LC filters with fractional bandwidths smaller than around 0.01 tend to become impractical, even when they can be realized, as mechanical and thermal stability of the components becomes an issue at such small bandwidths, and such filters are easily de-tuned by changes in temperature, or mechanical stresses.

Filters implemented using a single resonant circuit (resonator) do not have a very well-defined passband. The transfer function is not constant within the -3dB bandwidth of the filter, and it decays slowly outside of the passband. Thus, the bandwidth of the filter at the -30dB points is much wider than the bandwidth between the -3dB points. A measure of how closely a bandpass filter's passband approaches that of an ideal, rectangular filter, is the so-called *shape factor* of the filter response. Shape factor is defined as the ratio of the bandwidth at some large attenuation (say 30dB) and the bandwidth at the 3dB attenuation points. As a filter's transfer function approaches the ideal, rectangular shape, its shape factor approaches 1.0. A single-resonator LC filter is far from this ideal - for example, such a filter with a Q of 50 has a fractional bandwidth of $1/50$, or 0.02, and a shape factor (using the -30 dB and -3 dB bandwidths) of 31.6. Thus, the -30 dB bandwidth is more than 30 times as wide as the -3 dB bandwidth. Obviously, if a single resonator filter response is adjusted so that the signal of interest fills the -3dB bandwidth, an adjacent channel will not

be attenuated by very much. Thus, the single resonator filter is not appropriate for providing so-called *adjacent-channel rejection*. The shape of a filter's transfer function can be made to approach the idealized rectangular shape by using more (sometimes, *many* more) than a single resonator. For example, using only two resonators, a filter with fractional bandwidth of 0.02 and shape factor of 10.1 is easily obtained — this is more than a 3:1 improvement over the shape factor of a single-resonator filter. A comparison between the attenuation of a single-resonator filter and a 2-resonator filter, each having fractional bandwidth of 0.02, is shown in Figure 2.5.

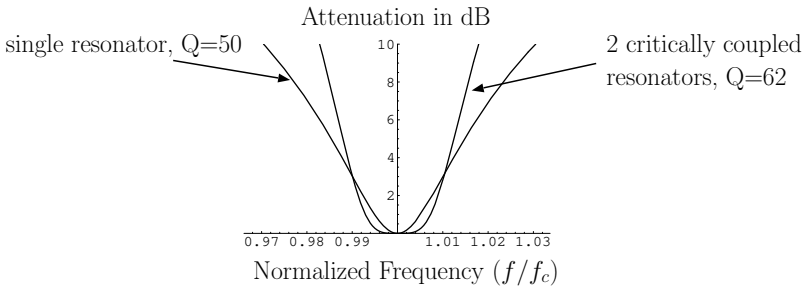


Figure 2.5: Attenuation vs normalized frequency for single and double resonator filters with the same fractional bandwidth. The single resonator filter has a shape factor of approximately 32, whereas the 2-resonator filter has a shape factor of 10.

2.2.1 Transmission-line and cavity resonator filters

For microwave applications it becomes feasible to utilize resonators based on distributed elements such as transmission lines, or waveguide cavities. Generally speaking, in their optimum implementations these resonators allow higher Q's than lumped LC resonators, and provide smaller minimum fractional bandwidths than provided by lumped LC filters. They are not easily tunable, however, restricting their application to fixed center frequencies and bandwidths. Filters based on distributed elements are most suitable for use as RF front-end filters in microwave receivers.

2.2.2 Filters based on piezoelectric devices - Quartz-Crystal Filter, Ceramic Filter

When very small fractional bandwidths and shape factors approaching 1.0 are required, it is necessary to turn to electro-mechanical resonators implemented using piezoelectric materials such as quartz crystals or ceramics, or electro-acoustic resonators implemented using surface-acoustic-wave (SAW) devices. These devices have Q-factors several orders of magnitude larger than those available from LC resonators, and they have much smaller thermal coefficients, allowing extremely narrowband filters to be realized. For example, a quartz crystal may have a Q of 100,000, allowing fractional bandwidths as small as 10^{-5} to be implemented. At a center frequency of 10 MHz, a fractional bandwidth of 10^{-5} corresponds to -3 dB bandwidth of 100 Hz. Unfortunately, the electro-mechanical resonators are useful

only at relatively low center frequencies (typically, < 30 MHz). Quartz-crystal and ceramic filters are fixed-frequency devices, and cannot be tuned over any appreciable range.

2.2.3 SAW filters

At frequencies in the range 30 MHz$\sim$2.5 GHz, filters based on surface-acoustic wave (SAW) devices provide a convenient means to realize filters with fractional bandwidths as small as 0.1% ($\Delta f/f_o = 0.001$) with excellent shape factors in a small package. These filters effectively fit many resonators within a small volume by converting an electrical signal into an acoustic signal, where the wavelength is 5 orders of magnitude smaller because of the slow speed of sound relative to the speed of electromagnetic wave propagation. By filtering the acoustic signal, it is possible to build filters which, effectively, contain 10's or hundreds of resonators within an extremely small package. This makes it possible to realize filters with excellent shape factors. SAW filters are fabricated for a fixed center frequency, and are not tunable.

2.2.4 Filter limitations dictate carrier-frequency conversion

Consider the fractional bandwidth required to make a single PCS-band CDMA (code-division-multiple-access) cellphone channel fill the -3dB bandwidth of a filter. Why fill the bandwidth of the filter with the signal of interest? Generally speaking, when a signal is accompanied by noise and interference, maximum signal-to-noise ratio will occur at the output of a filter when the filter shape is matched to the spectrum of the signal. The bandwidth of CDMA channel is 1.25 MHz, and the center frequency is approximately 1900 MHz. A fractional bandwidth of $1.25/1900 \approx .00066$ is required, with shape factor as close to 1 as possible, to select one channel while providing rejection of adjacent channels.

It is not possible to implement anything close to this small fractional bandwidth using practical filters based on any type of conventional electronic resonator (LC, transmission line, cavity resonators, etc). Quartz crystal filters can achieve the required fractional bandwidth, but not at a center frequency above 100 MHz or so. The solution is to move the carrier frequency of the desired channel to a lower frequency, without changing the bandwidth of the signal or distorting the modulation in any way. If the carrier frequency is lowered to, say 45 MHz, then the fractional bandwidth of the signal becomes $1.25/45 \approx 0.028$. This center frequency and fractional bandwidth are well within the range where SAW filters provide excellent performance. The concept of using *frequency conversion* to downconvert high frequency signal to a lower frequency where filtering and amplification is easier to achieve is embodied in the superheterodyne receiver.

2.3 The Superheterodyne Receiver

A block diagram of a single-conversion receiver is shown in Figure 2.6. The superheterodyne receiver operates by converting the input frequency of interest (f_c) to a fixed intermediate frequency (IF). This frequency conversion is performed by the mixer and local oscillator (LO). The intermediate frequency may be higher or lower than the carrier frequency of interest, but it is important to note that the IF is fixed (constant). Conversion to a fixed frequency takes advantage of the fact that it is much easier to realize narrow-band filters and stable, high-gain amplifiers if the frequency of operation doesn't change.

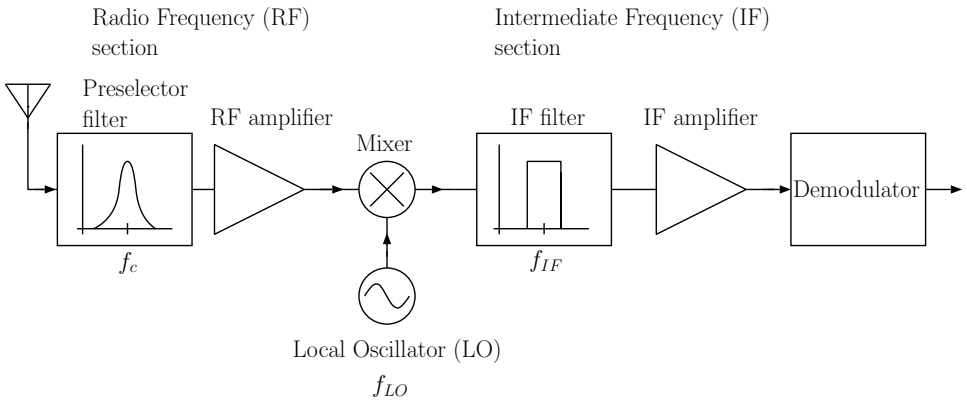


Figure 2.6: Superheterodyne receiver

A perfect multiplier can be used to model the operation of the mixer for preliminary discussions of the basic operation of the receiver, in which case the output of the mixer will consist of two signals with carrier frequencies equal to the sum and difference of the input and LO frequencies. Consider a modulated input signal with carrier frequency f_c which may be both angle- and amplitude-modulated in general:

$$s(t) = A(t)\cos[\omega_c t + \theta(t)] \tag{2.2}$$

The local oscillator signal will be an unmodulated carrier with frequency f_{LO}, i.e., $\cos(\omega_{LO}t)$. The output signal from the mixer will be

$$\begin{aligned} s_{out}(t) &= A(t)\cos[\omega_c t + \theta(t)]\cos(\omega_{LO}t) \qquad (2.3)\\ &= \frac{1}{2}\left\{A(t)\cos\left[(\omega_c - \omega_{LO})t + \theta(t)\right] + A(t)\cos\left[(\omega_c + \omega_{LO})t + \theta(t)\right]\right\} \end{aligned}$$

Note that the output from the mixer/LO consists of two signals with carrier frequencies $f_c - f_{LO}$ and $f_c + f_{LO}$. Note also that the amplitude and angle modulation that was present on the input signal has been transferred to the two output signals without any distortion.

The IF filter "picks out" one of the two signals and rejects the other. If a range of input frequencies is to be covered, then the LO will be tunable. Since the IF frequency is fixed, the LO frequency is adjusted, in practice, in order to make $f_{IF} = f_c + f_{LO}$ or $f_{IF} = |f_c - f_{LO}|$. The choice of whether the sum or difference frequency is picked out by the IF filter is determined in the design stage. If $f_{IF} < f_c$ the configuration is called *down-conversion*, since the carrier frequency has been converted down to f_{IF}. If $f_{IF} > f_c$, the receiver is said to employ *up-conversion*.

2.3.1 Image Frequencies

Just as there are two output signals from the mixer for each input frequency, there are two input frequencies to the mixer that will give an output at the IF frequency. In practice only one of these will be the desired frequency, and the other, undesired, frequency is called

the *image* frequency. If an undesired signal happens to have a carrier frequency that is the same as the image frequency, then that signal would be mixed into the IF filter's passband along with the desired signal. In the superhet (see Figure 2.6) the primary function of the preselector filter is to pass the desired signal and to reject signals at the image frequency. In a well designed receiver the image frequency is separated from the carrier frequency by a relatively large interval. Thus the preselector filter does not need to have a very narrow bandwidth. This is desirable, since the preselector filter often needs to be tunable. It is difficult to build a narrowband filter that has a tunable center frequency.

The preselector filter does not provide *adjacent channel* rejection, i.e., it does not reject signals that have carrier frequencies immediately adjacent to that of the desired signal. This is the function of the intermediate frequency (IF) filter which will determine the signal bandwidth of the receiver and, therefore, provide the rejection of adjacent channels. The bandwidth of the IF filter will normally be chosen to be just wide enough to accept the spectrum of the desired signal. Of course, interfering signals may still occur within the spectrum of the desired signal. In this case a "notch" filter can sometimes be used in the IF path to reject a very narrow slice of the signal spectrum which contains the interfering signal without significantly affecting the desired signal. The *adjacent channel* selectivity of a receiver is a function of the steepness at which the IF filter response rolls off outside of the passband.

2.3.2 Operation of the Mixer/LO Stage - Useful Relationships

For given IF and input frequencies, there are two possible LO frequencies that will cause the input frequency to be converted to the IF frequency:

$$f_{LO} = f_{IF} + f_C \quad \text{or} \quad f_{LO} = |f_{IF} - f_C| \tag{2.4}$$

For given IF and LO frequencies, there are two possible carrier frequencies that will give an output at the IF frequency:

$$f_{c1} = f_{LO} + f_{IF} \quad \text{and} \quad f_{c2} = |f_{LO} - f_{IF}| \tag{2.5}$$

One of these will be the desired carrier frequency, f_c, and the other will be an undesired image frequency, f_{IM}. The choice that was made for the LO frequency will determine which of these equations gives the image and/or desired frequency. The primary purpose of the preselector is to reject the undesired image frequency.

There are four possible "generic" configurations for a single-conversion superheterodyne receiver. Define f_{cmin} and f_{cmax} to be the lower and upper limits, respectively, of the input frequency range that the receiver is to cover. In practice the IF frequency will either be smaller than f_{cmin} or larger than f_{cmax}. These two cases can be further subdivided, because there are two possible choices for the LO tuning range for each choice of IF. The image frequency and the separation between the desired and image frequencies ($|f_{IM} - f_c|$) are summarized in Table 2.1 for each of the four cases.

2.3.3 Example - AM Broadcast Receiver.

A typical AM broadcast receiver covers the frequency range 540 - 1700 kHz. AM broadcast stations in the U.S. are assigned frequencies that are integer multiples of 10 kHz, therefore the adjacent channel separation is 10 kHz. The IF frequency is very often chosen to be 455

Table 2.1: Generic Configurations for a Single-conversion Superheterodyne Receiver

(1)	**Up - conversion:**	$\mathbf{f_{IF} > f_{cmax} > f_{cmin}}$			
(1a)	$\mathbf{f_{LO} = f_{IF} + f_c}$	$f_{IM} = f_c + 2f_{IF}$	$\lvert f_{IM} - f_c \rvert = 2f_{IF}$		
(1b)	$\mathbf{f_{LO} = f_{IF} - f_c}$	$f_{IM} = 2f_{IF} - f_c$	$\lvert f_{IM} - f_c \rvert = 2f_{LO}$		
(2)	**Down - conversion:**	$\mathbf{f_{IF} < f_{cmin} < f_{cmax}}$			
(2a)	$\mathbf{f_{LO} = f_{IF} + f_c}$	$f_{IM} = f_c + 2f_{IF}$	$\lvert f_{IM} - f_c \rvert = 2f_{IF}$		
(2b)	$\mathbf{f_{LO} = f_c - f_{IF}}$	$f_{IM} = \lvert f_c - 2f_{IF} \rvert$	$\lvert f_{IM} - f_c \rvert = 2f_{IF}$	if	$f_c > 2f_{IF}$
			$\lvert f_{IM} - f_c \rvert = 2f_{LO}$	if	$f_c < 2f_{IF}$

kHz. This is an example of "down-conversion." From (2a) and (2b), there are two choices for the local oscillator tuning range:

1. 995 kHz to 2155 kHz
2. 85 kHz to 1245 kHz

The "tuning ratio" ($R = f_{LOmax}/f_{LOmin}$) is an important factor in determining the cost and stability of an oscillator. It is easier to build a stable oscillator if the tuning ratio is small. Note that the tuning ratio for choices (1) and (2) are 2.16 and 14.6, respectively. For this reason the first choice given above is almost always used. In fact, in most (but not all) cases the higher of the two possible LO frequencies will be used (i.e., (1a) or (2a) from Table 2.1), because this leads to the smaller tuning ratio. In this particular example the second choice has another disadvantage, since it would require the local oscillator to tune through the IF frequency, i.e., to receive a signal at 910 kHz; the second choice requires the local oscillator to be tuned to 455 kHz. It is very likely that some of the local oscillator signal would leak into the IF stage of the receiver. The result would be interference with the desired signal, as well as possible overload of the IF and succeeding stages.

Assuming the LO tunes from 995 to 2155 kHz, then the image frequency will always be (from (2a)) given by $f_{IM} = f_c + 2f_{IF}$. For example, if we wish to tune the receiver to WILL at 580 kHz, the preselector filter would be tuned to 580 kHz and the local oscillator would be tuned to 1035 kHz. The bandwidth of the preselector needs to be narrow enough to "reject" any incoming signals at the image frequency $f_{IM} = 1490$ kHz. The IF filter needs to be 10 kHz wide in order to provide adjacent channel selectivity.

As the radio is tuned across the band, it is necessary to tune both the LO and the preselector simultaneously. This is why the tuning knob on mechanically tuned AM radios is connected to a two-section, or "ganged," variable capacitor. One section tunes the LO, and the other tunes the preselector filter. One or more adjustments is usually provided to insure that the two sections "track" each other.

Note carefully that for carrier frequencies less than 1245 kHz, the local oscillator frequency will be less than 1700 kHz and therefore inside the AM broadcast band. It is sometimes possible to hear this LO signal on another nearby AM radio. This phenomenon is undesirable and is caused by the local oscillator signal being fed back through the mixer and out the antenna, or by direct radiation from the LO circuitry. Commercial and military receivers are carefully designed to minimize such radiation. This is done by using balanced mixers that have very low direct output-to-input coupling coefficients and by carefully shielding the LO circuitry to avoid direct radiation. If the LO frequency is well removed from the carrier frequency, a notch filter can be put before the mixer to reduce the LO signal that is

coupled to the antenna.

The configuration of a standard FM broadcast receiver is described in the next example.

2.3.4 Example - FM Broadcast Receiver

The FM broadcast band covers 88 - 108 MHz. The channels are separated by 200 kHz and are assigned to odd multiples of 100 kHz. In almost all cases the IF frequency is chosen to be 10.7 MHz. As in the previous example "high-side" LO is often used, i.e., the LO tunes from 98.7 to 118.7 MHz.

Notice that the IF used for FM broadcast receivers is substantially higher than that used for AM receivers. This choice is motivated by the fact that the image frequency is separated from the desired carrier frequency by $2f_{IF}$. Generally speaking, in order to make the preselector filter relatively easy to build and tune, the IF frequency must be raised as the carrier frequency increases.

2.3.5 Up-conversion versus Down-conversion

The AM and FM broadcast receivers considered above both use down-conversion and the higher of the two possible LO frequencies. The image frequency was $2f_{IF}$ above the desired carrier frequency. In the AM receiver case the width of the carrier frequency band of interest is larger than $2f_{IF}$. This means that the image frequency can fall within the band of interest, thus necessitating the use of a tunable preselector filter that has a bandpass characteristic. In receivers designed to cover very wide frequency bands the use of a tunable preselector becomes impractical, because it is difficult to build filters whose center frequency must be tuned over a wide frequency range. In this case it is advantageous to use up-conversion, because with up-conversion the image frequency will always be larger than f_{cmax}. Thus, the image frequency will always fall outside (above) the band of interest, and a preselector filter with a lowpass characteristic is sufficient. A more practical choice, however, would be a fix-tuned bandpass filter that passes the entire carrier frequency band of interest. The bandpass filter is usually the best option, since it minimizes the possibility of interference and receiver overload from strong signals outside of the frequency range of interest. The up-conversion scheme can be a significant advantage, especially in a receiver that is designed to cover a wide frequency range, since only the LO needs to be tuned.

2.3.6 Single- versus Double-conversion

There is a lower limit to the fractional bandwidth ($\Delta f/f_o$) that can be realized with practical filters. This sets an upper limit on the intermediate frequency for a given signal spectrum bandwidth. However, because the separation of the image frequency from the desired frequency is usually twice the intermediate frequency, a large intermediate frequency is desired. A large separation is desirable because it makes rejection of the image easier. In some cases these conflicting considerations make it impossible to select an intermediate frequency that is low enough to achieve adequate adjacent channel selectivity but large enough to make it possible to reject the image frequency with a simple preselector filter. Then it becomes necessary to use a double-conversion scheme (Figure 2.7) where the carrier frequency of interest is first converted to a relatively high IF (allowing rejection of the image frequency with a simple preselector filter) and then converted again to a lower IF, in order to obtain the

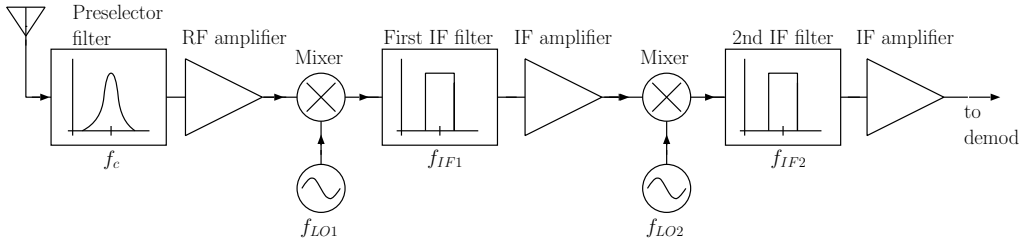


Figure 2.7: Double-conversion superheterodyne receiver. The first conversion ($f_c \rightarrow f_{IF1}$) can be either up- or down-conversion. The second conversion ($f_{IF1} \rightarrow f_{IF2}$) is always a downconversion.

required bandwidth and adjacent channel selectivity. In practice, the first conversion can be either an up- or down-conversion, but the second conversion is always a down-conversion.

In certain cases there are advantages to using more than two conversions, i.e., triple-conversion, although problems with spurious responses multiply rapidly as the number of frequency conversions increases. The image frequency causes one type of spurious response that is commonly observed in inexpensive AM and FM broadcast receivers where a certain radio station is received when the receiver is tuned to a frequency other than the true carrier frequency of the signal. There are several other origins for spurious responses involving both external signals and signals generated within the receiver. These will be discussed in some detail in Chapter 12. An important step in the design procedure is to identify and evaluate the impact of spurious responses on the overall operation of the receiver.

In a multiple-conversion receiver the bandwidth of the last IF stage sets the overall bandwidth of the receiver. In a double-conversion receiver the bandwidth of the first IF filter would be chosen to reject "secondary" image frequencies within the first IF bandwidth. A secondary image would be an undesired frequency within the passband of the first IF filter that could be mixed into the second IF filter's passband. The second IF filter's bandwidth would usually be just wide enough to pass the entire spectrum of the desired signal.

2.3.7 The 1/2-IF response

Nonlinearity in the RF, LO, and mixer sections of the receiver causes mixing products involving harmonics of the RF and LO signals to be present at the mixer output. This causes spurious responses, i.e. the receiver responds to input frequencies other than the desired frequency. One example of this type of spurious response is the so-called *1/2-IF response*. The 1/2-IF response results from the second harmonic of an input signal mixing with the second harmonic of the LO signal to produce an output at f_{IF}. The mechanism is illustrated here for the case of high-LO and down-conversion. Suppose that the receiver is tuned to receive a desired signal with carrier frequency f_C. The LO will be tuned to

$$f_{LO} = f_C + f_{IF} \tag{2.6}$$

Suppose a signal with frequency $f_S > f_C$ is also present at the receiver input. If the second harmonic of this signal mixes with the second harmonic of the LO, the mixer output will have a component at $|2f_S - 2f_{LO}|$. This signal will interfere with the desired signal if

$|2f_S - 2f_{LO}| = f_{IF}$, or if

$$f_S = f_C + f_{IF} \pm f_{IF}/2.$$

Thus a spurious response occurs when $f_S = f_C + (3/2)f_{IF}$ or $f_S = f_C + f_{IF}/2$. Notice that both of these responses are "closer" to the desired signal than is the image frequency. The closest spurious response is separated from the desired signal by 1/2 of the IF frequency.

The 1/2-IF response plays a role in the choice of f_{IF}. Suppose that the signals of interest are located within a carrier frequency range denoted by R. It is desirable to use a fix-tuned preselector that passes the entire carrier frequency band of interest. In order for the preselector to reject potential spurious responses it is necessary for all significant spurious responses to fall outside of the carrier frequency range of interest. If only the image response is considered, then this constraint is satisfied if f_{IF} is chosen such that $2f_{IF} > R$, or $f_{IF} > R/2$. When the 1/2-IF response is considered, then spurious responses will fall outside of the desired range if $f_{IF}/2 > R$, or $f_{IF} > 2R$. Notice that when the 1/2-IF response is considered the minimum value of f_{IF} is 4 times as large as when only the image response is considered.

2.4 Zero-IF receiver

In many applications, it is highly desirable to be able to implement an entire receiver in an integrated circuit, without the need for bulky and expensive IF filters, which require that signals be routed off-chip to the filter. For such applications the zero-IF (ZIF) (also called direct-conversion or homodyne) approach has gained popularity. There are various ways of explaining the evolution of this concept, e.g. it is sometimes described as the limit in which the intermediate frequency of a superhet approaches zero. Perhaps it is more accurate to say that a ZIF receiver amounts to using a quadrature demodulator *as the receiver*. A ZIF receiver is shown in Figure 2.8. The difference between this and a basic quadrature

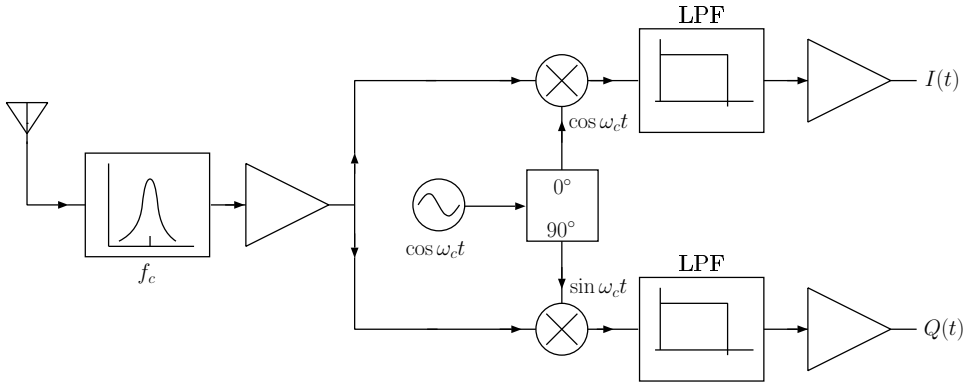


Figure 2.8: Direct-conversion receiver. Depending on the application, the local carrier oscillator may need to be derived from a carrier recovery circuit.

demodulator is the addition of an RF filter and low-noise amplifier (LNA) in front of the demodulator, and addition of baseband amplifiers after the lowpass filters in each arm of the demodulator. In a ZIF receiver, the filters responsible for selecting a desired channel,

and rejecting undesired channels, are the lowpass filters at the output of each multiplier (mixer). This is attractive, since active lowpass filters can be applied here. A very flexible receiver architecture results if the I and Q outputs of the ZIF demodulator are sampled with an analog-to-digital converter so that digital signal processing techniques can be applied to implement very good channel selection filters and to perform the signal processing required for demodulation. This relaxes the requirement on the analog LPFs to the simpler task of providing the anti-aliasing function.

The ZIF concept looks deceptively simple. In practice, a number of issues combine to make it a challenge to obtain good performance using this architecture. In a ZIF receiver, most of the gain will be provided by the baseband amplifiers after the LPFs. Hence, desired signals at the output of the multipliers are generally much smaller than they would be if the quadrature demodulator was used after the IF stage of a conventional superhet receiver, where signal levels are relatively high by virtue of the high IF gain. The small desired signals are easily corrupted by noise and DC offsets contributed by the mixers, local oscillator, and the DC-coupled baseband amplifiers. The local oscillator contributes to the DC offset problem because self-mixing of the local oscillator with itself results in a DC component. In addition, the self-mixing brings the phase noise sidebands of the local oscillator down into the baseband spectrum, especially in the region near DC. The downconverted LO phase noise adds to 1/f noise (flicker noise) contributed by the mixer and amplifiers, and can significantly degrade the signal to noise ratio for signal components located near the carrier frequency, which are converted to very low frequencies near DC after the mixers.

Development of ZIF receivers is an active research area, with efforts focused on minimizing mixer and LNA noise, development of extremely linear LNAs, and development of low phase noise oscillators.

Some of the advantages of the zero-IF configuration can be obtained while avoiding the problems associated with DC offsets and 1/f noise near 0 Hz by downconverting the signal to a "low-IF" instead of all the way to 0 Hz. Figure 2.9 compares the architectures of a conventional superhet receiver, the low-IF, and the zero-IF configurations.

2.5 Software Defined Radio

The term Software Defined Radio (SDR) has become popular to describe transmitter and receiver systems that utilize digital signal processing techniques to minimize the amount of specialized analog hardware required for a specific application, and to allow reconfiguring the communications system with changes in software alone. This is especially important in view of the large number of standards in place, and under development, for the various communications applications; cellular telephony is a good example - as of today, there are at least 5 popular standards for modulation format in use, implemented in two widely separated frequency bands.

Considering the receiver portion of a SDR-based transceiver, the "ideal" SDR would employ an A/D converter to sample the voltage across the antenna terminals, and then perform all filtering and demodulation in software. This is not practical at the present time, mainly due to the limitations on A/D converter dynamic range (number of useful bits of resolution) and sampling speed. Practical implementations of SDR's, in order of increasing amount of software control and decreasing amount of hardware required, include:

1. sample the I and Q outputs of a quadrature demodulator, allowing fine-tuning, carrier synchronization, demodulation, to be carried out entirely in software. The analog

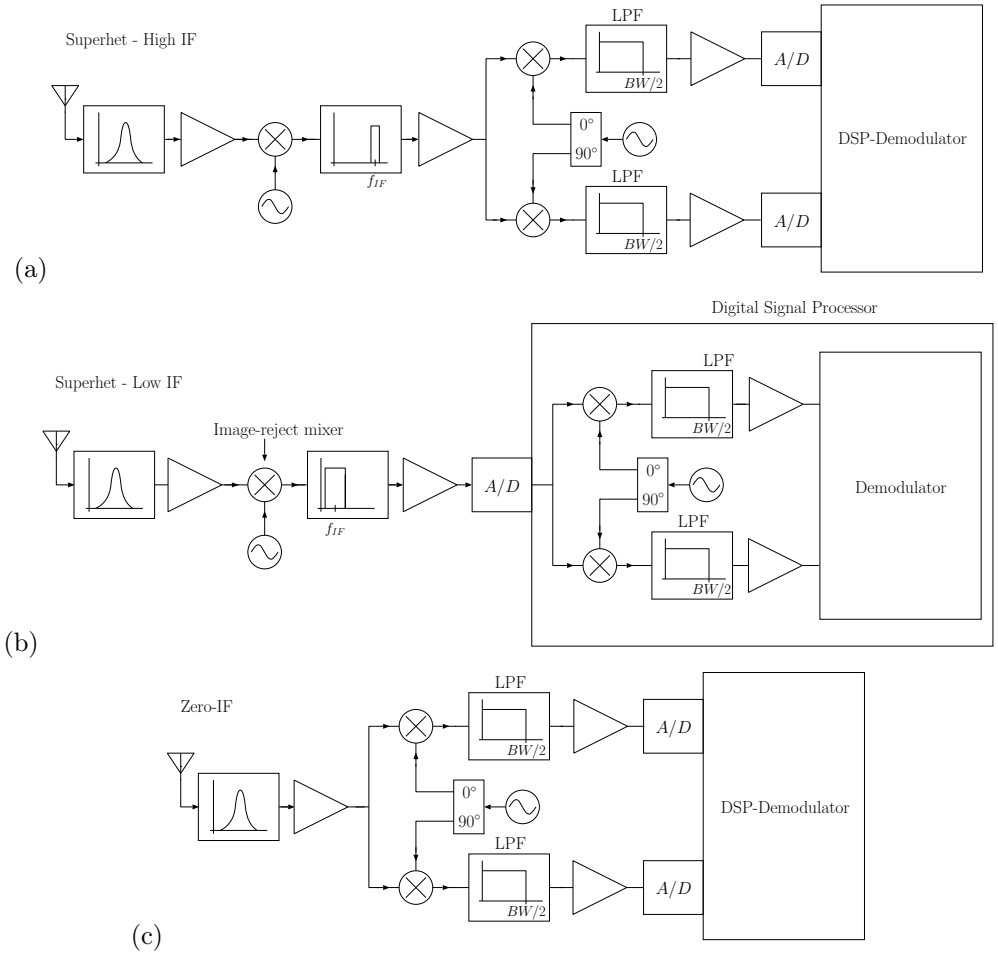


Figure 2.9: Comparison of superheterodyne receivers implemented with (a) relatively large IF, (b) relatively low IF, and (c) zero-IF. Each configuration includes analog-to-digital (A/D) conversion and digital signal processing stages. Use of a low IF in (b) makes it possible to digitize the IF output directly, allowing the quadrature downconversion and demodulation stages to be implemented in a DSP. The need for only one A/D converter is an advantage for the low-IF superhet receiver. On the other hand, the preselector in a low-IF receiver will not be able to provide much image rejection because the image response will be located close to the desired frequency. An image-reject mixer can be used to provide additional image rejection (see homework problem 18). The advantage of having only a single A/D converter comes at the cost of the additional complexity of the image-reject mixer.

frontend must implement some filtering, amplification, quadrature LO generation, and multiplication/mixing of the RF and LO signals.

2. when very high speed A/D converters are available, a wide-bandwidth slice of the filtered/amplified antenna output, or a wideband IF output, can be sampled at high speed and digitally downconverted to produce sampled I and Q components with narrower bandwidth and lower sampling frequency using a high speed dedicated signal processor. Several specialized *digital downconverter* chips are available to perform this function. Typically, the digital downverter samples at high rate and then applies a rudimentary digital filter to the samples. The main purpose of the digital filter is to place nulls at frequencies that will be folded into the desired slice of the spectrum after decimation is performed. After the filter, the high-speed samples are decimated to a rate that is at least twice the bandwidth required for subsequent processing. The lower-rate samples are then processed by a conventional programmable DSP or computer. The analog frontend provides filtering, amplification, and possible downconversion to an IF; generation of the quadrature LO, mixing to I/Q channels, and demodulation is performed in the digital downconverter and in software.

For those interested in learning more about SDR techniques through experimentation, there are two excellent open-source programs available for *NIX platforms:

1. Linrad (www.nitehawk.com/sm5bsz/linuxdsp/linroot.htm) is a software receiver for *NIX plaftorms that uses a stereo sound card to sample the I and Q outputs of a quadrature demodulator. The program has AM/FM/SSB demodulators, and provides a realtime waterfall display. This program is intended for demodulation of weak, narrowband signals. A great deal of work has been put into implementing state-of-the-art noise blanking, making this the program of choice for amateur radio operators experimenting with communications using signals bounced off of the moon; called earth-moon-earth, or eme, communications.

2. GnuRadio is the GNU software radio (www.gnu.org/software/gnuradio/index.html). According to their web site: “GNU Radio is a collection of software that when combined with minimal hardware, allows the construction of radios where the actual waveforms transmitted and received are defined by software. What this means is that it turns the digital modulation schemes used in today’s high performance wireless devices into software problems.”. GnuRadio software runs on *NIX, and can use various input devices, including sound cards and high-speed A/D cards. Many state-of-the-art DSP techniques applicable to SDR are available as modules. Receiving systems can be easily put together and reconfigured using a Python interface.

2.6 References

1. *Communications Receivers: Principles and Design*, Ulrich L. Rohde, Jerry Whitaker, and T. T. N. Bucher, 2nd Edition, McGraw Hill, New York, 1997.

2. *Single-Sideband Systems and Circuits*, Edited by William E. Sabin and Edgar O. Schoenike, McGraw Hill, New York, 1987.

3. *The Science of Radio*, Paul J. Nahim, 2nd Edition, Springer Verlag, New York, 2001.

4. *The Design of CMOS Radio-Frequency Integrated Circuits*, Thomas H. Lee, Cambridge University Press, 1998.

5. *RF Microelectronics*, Behzad Razavi, Prentice Hall, 1998.

2.7 Homework Problems

1. The transfer function for a regenerative amplifier is found to be

$$T(\omega) = \frac{GF(\omega)}{1 - AGF(\omega)} \tag{2.7}$$

where A and G are assumed to be positive real constants with AG < 1. Suppose that the filter frequency response function is approximated by a Gaussian function, i.e.,

$$F(\omega) = \exp[-(\omega - \omega_c)^2/B^2] \tag{2.8}$$

This filter has a 3 dB bandwidth W = 1.177 B.

   (a) Find an expression for the 3 dB bandwidth of the regenerative amplifier's transfer function, T(ω).

   (b) Evaluate the 3 dB bandwidth for the cases AG = 0.9 and AG = 0.99.

2. The transfer function for a regenerative amplifier is:

$$T(\omega) = \frac{GF(\omega)}{1 - AGF(\omega)}$$

where A and G are positive real constants with $AG < 1$. Suppose that the filter is implemented with a single series or parallel RLC resonant circuit, in which case the frequency response function of the filter, $F(\omega)$, will have the form:

$$F(\omega) = \frac{1}{1 + jQ(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})},$$

where the factor Q is a dimensionless constant that controls the bandwidth of the filter.

   (a) Find an expression for the bandwidth of $F(\omega)$ at attenuation level α decibels — the "-α dB bandwidth" — where α is the attenuation expressed in dB. Denote this bandwidth by $\Delta f_{-\alpha}$ and express your result in terms of the power ratio $r = 10^{\alpha/10}$, the parameter Q, and the center frequency $f_o = \omega_o/(2\pi)$. Note that $|F(2\pi f)|$ is a bandpass function with peak at $f = f_o$. The $-\alpha$ dB bandwidth is defined to be the separation between the two frequencies f_1 and f_2 which satisfy $f_1 > 0$, $f_2 > 0$ and $|F(2\pi f_1)| = |F(2\pi f_2)| = |F(2\pi f_o)|/\sqrt{r}$.

   (b) Using your answer to part a, write down the expression for the -3 dB bandwidth of $F(\omega)$ and make note of it. We will use this result often during the rest of the semester.

   (c) Find an expression for the -3 dB bandwidth of $T(\omega)$. Make sure that your result reduces to the result found in part b when $A \to 0$. Discuss how the magnitude of the loop gain (AG) affects the peak gain of $T(\omega)$ and the bandwidth of $T(\omega)$.

   (d) Refer to your results from part c, and comment on how the product of the peak gain and bandwidth of $T(\omega)$ (the *gain-bandwidth product*) depends on the feedback gain parameter, A.

(e) Find an expression for the *shape factor* (SF) of the the regenerative amplifier's frequency response, $T(\omega)$, where

$$SF = \frac{\Delta f_{-30\text{dB}}}{\Delta f_{-3\text{dB}}}.$$

Your result will depend on the loop gain, AG. Note that when the feedback parameter $A = 0$, then $T(\omega) = GF(\omega)$, and your result will reduce to the shape factor of $F(\omega)$. Find the numerical value of the shape factor for $AG = 0$, $AG = 0.5$, and $AG = 0.99$. (Note - a good channel selection filter will have a shape factor smaller than 2. A really good filter might have shape factor of 1.2.)

3. Consider a bandpass filter consisting of a number, N, of cascaded filters which are isolated from each other by buffer amplifiers. Assume that each filter has the transfer function given by $F(\omega)$ in problem 2. The overall transfer function of the system will be $F(\omega)^N$. Derive an expression for the -3 dB bandwidth of the cascade of N filters. Express the bandwidth as a product of two terms: (i) the bandwidth of a single filter (this was found in problem 2, part a), and (ii) a bandwidth reduction factor which depends only on N. Calculate the bandwidth reduction factor for $N = 2, 3, 4, 5, 6$.

4. Suppose we want to design a superhet receiver to cover the frequency range 1 - 30 MHz. Two possible intermediate frequencies (IF's) for the receiver are:

(i) 500 kHz
(ii) 60 MHz

For these IF's:

(a) Find the possible local oscillator frequencies (f_{LO}) and image frequencies (f_{IM}) for reception of signals with carrier frequencies of 1, 5, 15, and 25 MHz. Make a table.

(b) What are the LO tuning ratios required to cover the entire 1-30 MHz range for each possible choice of LO frequency?

(c) From the standpoint of the easiest and least expensive preselector and LO design, what is the optimum configuration? Assume that the required IF bandwidth is easy to obtain, whatever the choice of the LO frequency. For the optimum configuration, describe the required preselector frequency response. Is it necessary to tune the preselector?

5. Consider a single conversion superhet receiver that uses "low-LO" and downconversion.

(a) If the signals of interest have carrier frequencies in the range 140-170 MHz and $f_{IF} = 21.4$ MHz, find the tuning range for the local oscillator. Specify the minimum and maximum local oscillator frequencies.

(b) If the receiver is tuned to receive a desired signal with carrier frequency $f_c = 150$ MHz, find the image frequency.

(c) Assume that the preselector will have an ideal rectangular frequency response function. Does the preselector for this receiver have to be tunable? (Yes or no). Explain your answer. Do not consider the 1/2-IF response for this question.

(d) When the receiver is tuned to receive a desired signal with carrier frequency $f_c = 150$ MHz, denote the frequency of a signal that could cause interference if the second harmonic of the interferer mixes with the second harmonic of the LO by f_i. Find all possible values of f_i.

6. Consider the double-conversion receiver shown in Figure 2.10.

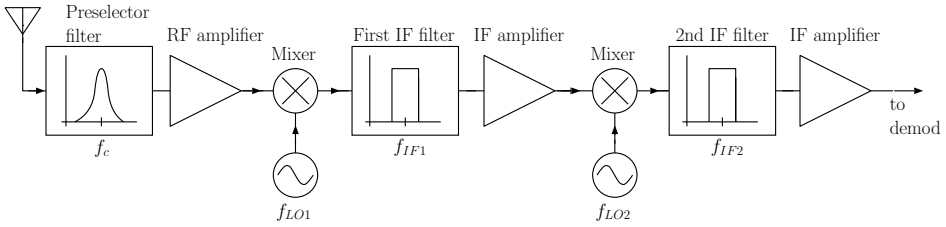


Figure 2.10: Double-conversion receiver

Note that a double-conversion receiver requires two IF filters. Assume that the receiver is to cover the 2-to-30 MHz range. The center frequency of the first IF filter is 50 MHz and that of the second is 500 kHz.

(a) Specify the possible frequencies (f_{LO1} and f_{LO2}) for a receiver covering 2 to 30 MHz. You must specify a range of frequencies for f_{LO1}, and a single frequency for f_{LO2}. Be sure to list both possibilities for each.

(b) Assume the first IF filter is a bandpass filter with ideal (rectangular) frequency response. What is the maximum bandwidth of the filter if secondary images are to be avoided? A secondary image is an undesired frequency within the first IF filter's passband that could be mixed into the second IF filters passband. Assume that the second IF filter has a rectangular response with bandwidth 10 kHz.

7. We will design a receiver that can receive lower sideband signals. A block diagram of the receiver is shown in Figure 2.11. Note that the SSB signal is demodulated using a mixer and beat-frequency oscillator (BFO). The BFO and second mixer can be thought of as a second frequency conversion stage where the "second IF" is at 0 Hz (DC). The purpose of this stage is to translate the signal spectrum down to base-band, i.e., to translate the carrier frequency to DC.

Suppose the input signal is an LSB signal with carrier frequency f_c =14.3 MHz and bandwidth = 3 kHz as shown in Figure 2.12. You may assume that the IF filter has an ideal rectangular bandpass characteristic with center frequency of 9.0 MHz and bandwidth BW_1= 3 kHz. *Thus, the IF filter bandwidth is just wide enough to pass the LSB signal's spectrum.*

(a) Give two choices for the frequency of the first local oscillator, f_{LO1}. Be careful: make sure that your choice of local oscillator frequency will cause the entire spectrum of the desired signal to fall within the IF filter's passband.

(b) For each choice given in part 7a, sketch the spectrum of the signal as it would appear just after the IF filter.

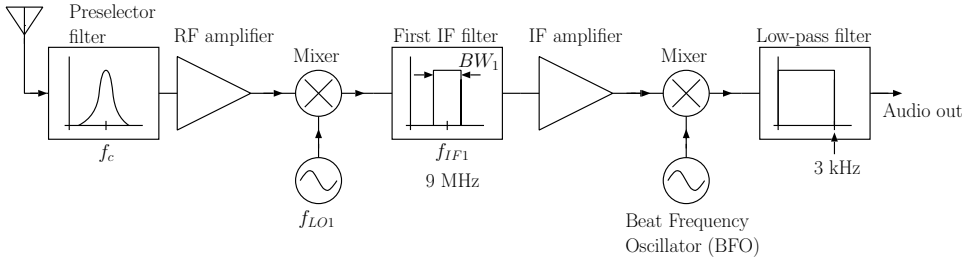


Figure 2.11: Receiver to receive lower sidebands

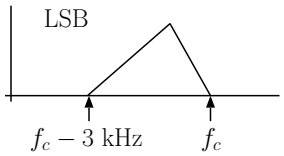


Figure 2.12: LSB signal

(c) For each choice given in part 7a, give the band of frequencies that correspond to the image response of the receiver.

(d) For each choice given in part 7a, give the frequency of the beat-frequency oscillator such that the signal is properly demodulated.

8. Consider the triple-conversion receiver shown in Figure 2.13. Suppose that $f_{LO1} =$ 1300 MHz, $f_{LO2} = 1410$ MHz, and $f_{LO3} = 179.3$ MHz. Note that the only filter is the final IF filter which has an ideal rectangular frequency response function centered at f_{IF} =10.7 MHz:

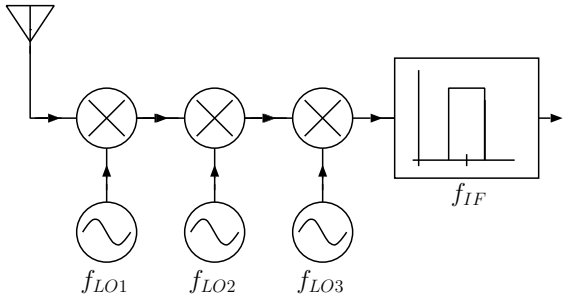


Figure 2.13: Triple-conversion receiver

(a) Find and list all input frequencies that will give an output from the final IF filter. You may ignore the finite bandwidth of the IF filter, i.e., list only the input frequencies that will give an output at exactly 10.7 MHz.

(b) How many input frequencies would be on your list if there were 4 conversions instead of 3?

9. Consider the design of the receiver portion of a portable cellular telephone. The cellular telephone receives frequency-modulated that are transmitted from a base-station. The received signals have a bandwidth of 30 kHz and a carrier frequency, f_c, somewhere within the frequency range 869-894 MHz. While receiving, the telephone must simultaneously transmit frequency-modulated signals back to the base-station at a carrier frequency f_c-45 MHz, i.e., the telephone transmits at a frequency 45 MHz below the frequency at which it receives.

   Sketch a block diagram of a double-conversion superhet receiver that could be used for receiving cellular phone signals. Your design should have a tunable first local oscillator that will serve a dual purpose. This oscillator will also serve as the carrier oscillator for the signals transmitted by the telephone. Assume that the second IF filter has a center frequency of 5.5 MHz.

   Specify the following and state the reasoning that led to your choice.

   (a) The center frequency of the first IF filter.

   (b) The maximum bandwidth of the first IF filter. Assume that the filter response is symmetric around the center frequency. Indicate what considerations led to your choice for the maximum bandwidth.

   (c) The bandwidth of the second IF filter.

   (d) Sketch an appropriate transfer function for the preselector.

   (e) Specify the tuning range for the first LO, i.e., specify the minimum and maximum first LO frequencies.

   (f) Specify the image frequency when the phone is set up to receive signals at f_c=875 MHz.

10. Design a double-conversion receiver (refer to Fig. 2.10) that will receive frequencies in the range 144-148 MHz. The signal bandwidth is 10 kHz. The first IF frequency is to be 10.7 MHz. The second IF frequency is 455 kHz. You also know that the receiver is going to be operated in a region where very strong local stations exist in the frequency range 122-125 MHz.

    Specify the following:

    (a) The bandwidths of the first and second IF filters, BW_1 and BW_2. Explain your choices.

    (b) The frequency (or frequency range) for each local oscillator, f_{LO1} and f_{LO2}. Explain your choices. "Tuning ratio" is not a significant consideration in this receiver, so do not invoke "smaller tuning ratio" as a reason for preferring one choice for f_{LO1} over another one.

    (c) The preselector filter would normally be a bandpass filter in this application. Is it necessary to use a tunable (variable) preselector filter? Why or why not?

11. Consider a single-conversion superhet receiver with $f_{IF} = 260$ MHz. The receiver must be able to tune to Sirius and XM satellite radio signals. These signals have

carrier frequency 2332.5 MHz and bandwidth 25 MHz. The receiver must present the entire 25 MHz bandwidth to a downstream signal processor that is responsible for demultiplexing the many audio channels that are present on the received signal. Thus, the IF filter will have center frequency 260 MHz, and bandwidth 25 MHz.

(a) Specify two possible frequencies for the local oscillator.

(b) For each answer in part (a.), specify the image frequency.

(c) Suppose that the receiver is implemented using "low-LO". Give the frequency of a signal that would cause interference as a result of the "1/2 IF response" mechanism.

12. Consider the double-conversion superhet receiver in Figure 2.14:

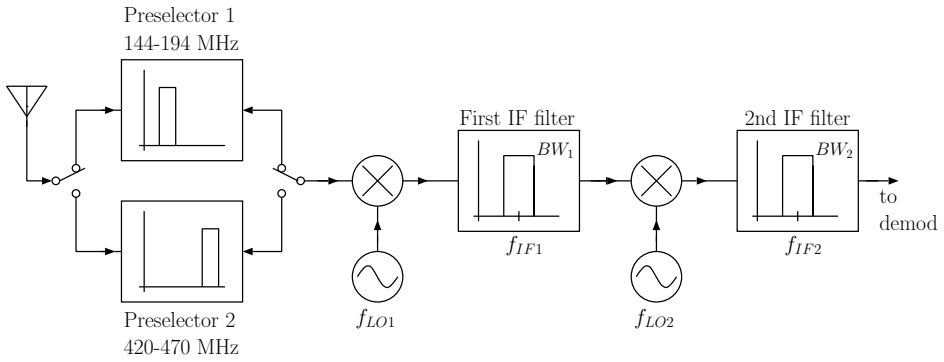


Figure 2.14: Dual band superhet receiver

This receiver is to be designed to cover two different carrier-frequency bands (144-194 MHz and 420-470 MHz) with a single tunable first local oscillator and first IF Filter. To change the carrier-frequency band, only the preselector filter will be switched as shown in Figure 2.14. You may assume that the signals of interest have a bandwidth of 15 kHz.

(a) Specify the first intermediate frequency (f_{IF1}). There are two possibilities - find both of them.

(b) For each possible IF found in part (a), specify the 50 MHz-wide range of frequencies covered by the first local oscillator (f_{LO1}).

(c) Specify the smallest and the largest bandwidth (BW_1) that could be used for the first IF filter. Explain what factors determined your choices. You may assume that the second IF filter has an ideal rectangular frequency response function with $f_{IF2} = 10.7$ MHz and $BW_2 = 15$ kHz.

13. The double conversion receiver shown in Figure 2.15 uses a special system to generate the first local oscillator signal. This is called the "Wadley Loop" drift cancelling system. It is used to make the receiver tuning insensitive to relatively small changes in the

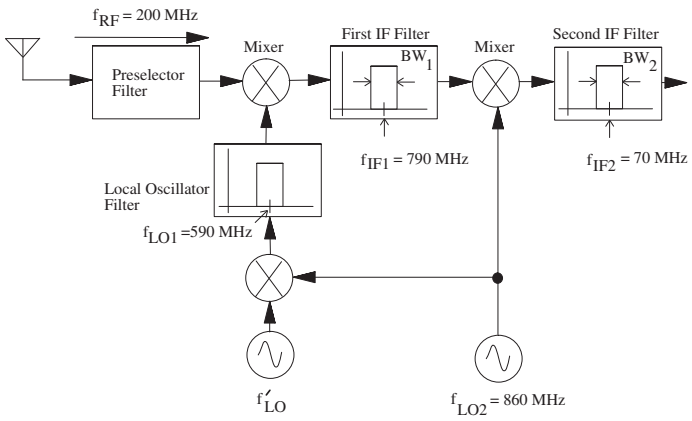


Figure 2.15: "Wadley Loop" drift canceling system

second local oscillator frequency that arise because of frequency drift. This makes it possible to use a relatively inexpensive oscillator for the second LO.

Suppose the input signal of interest has carrier frequency f_{RF}=200 MHz. The first IF filter is centered at f_{IF1}=790 MHz and the second IF filter is centered at f_{IF2}=70 MHz. The local oscillator filter is centered on 590 MHz. The RF signal has a bandwidth of 50 kHz. Assume that the IF filters have ideal, symmetrical rectangular passbands with bandwidth BW_1 and BW_2 as shown in Figure 2.15.

(a) Specify the bandwidth of the second IF filter, BW_2. State the reason for your choice.

(b) Give the maximum allowable bandwidth of the first IF filter (BW_1) so that secondary images will be rejected.

(c) Determine f'_{LO} such that an input signal with carrier frequency of 200 MHz will be centered in the passband of both the first and second IF filters when f_{LO2}=860 MHz and such that the frequency of the desired signal at the output of the second mixer is independent of the exact value of f_{LO2}. Hint: Write f_{LO2} as f_{LO2}=860 + δ and find the choice for f'_{LO} that makes the frequency after the second mixer independent of δ. There is only one correct choice for f'_{LO}.

(d) If you have done part 13c correctly, then the frequency of the desired signal after the second mixer will not change if f_{LO2} drifts away from its nominal value of 860 MHz. This does not mean that f_{LO2} can take on any value, however. Suppose that δ represents the frequency drift, i.e., suppose that f_{LO2}=860 + δ. What factor(s) limit the allowable size of δ?

14. Consider a single conversion superhet receiver with $f_{IF} = 70$ MHz. The receiver must be able to tune to carrier frequencies in the range 1930 MHz $\leq f_{RF} \leq$ 1975 MHz.

(a) Specify the two possible frequency ranges for the first local oscillator.

(b) For each answer in part (a.), specify the image frequency when the receiver is tuned to receive a desired signal with carrier frequency $f_{RF} = 1940$MHz.

(c) In one or two sentences, explain how the "1/2 IF response" of a receiver comes about.

(d) Suppose that the receiver described in part (a.) is implemented using high-LO. Give the frequency of a signal that would cause interference as a result of the "1/2 IF response" mechanism when the receiver is tuned to receive a desired signal with carrier frequency $f_{RF} = 1935$ MHz.

15. You are designing a superhet receiver for a PCS-band cellular phone handset to receive carrier frequencies in the range $1930\,\text{MHz} \leq f_{RF} \leq 1975$ MHz. Assume that the receiver will employ high-LO.

(a) Suppose that the preselector has a fixed rectangular bandpass response that passes 1930-1975 MHz. Determine the smallest value for the IF that will allow the preselector to reject signal(s) at the image frequency.

(b) Suppose that the preselector is the same as described in part a. Determine the smallest value for the IF that will allow the preselector to reject signal(s) whose second harmonic could mix with the second harmonic of the LO to produce an output at the IF.

16. Consider a double conversion superhet receiver with first IF and last IF denoted by f_{IF1} and f_{IF2}, and bandwidth of the first and last IF filters denoted by Δf_1 and Δf_2, respectively. You may assume that both IF filters have ideal rectangular response functions that are symmetrical about the center frequencies and that $f_{IF2} \ll f_{IF1}$.

(a) Give an expression for the maximum bandwidth of the first IF filter, $\Delta f_{1,max}$, such that secondary images will be rejected. You do not need to consider the half-IF response.

(b) Suppose, for the purpose of this problem, that it is impractical to build or acquire IF filters with fractional bandwidth $\Delta f / f_{IF}$ smaller than 2%. For a receiver using a last IF of 455kHz, specify the largest practical value for the first IF. For this calculation, you may assume that the bandwidth of the second IF filter (Δf_2) is small enough to ignore.

17. Consider a single conversion superhet receiver that uses downconversion to an IF denoted by f_{IF} and "high-LO". We showed that when such a receiver is tuned to receive a desired signal with carrier frequency f_c, an undesired signal with frequency $f_i = f_c + \frac{f_{IF}}{2}$ could cause interference if the second harmonic of the undesired signal mixes with the second harmonic of the LO. This is called the "half-IF" response because the potential interferer is offset from the desired carrier frequency by $\frac{f_{IF}}{2}$. Find the offset from the desired carrier frequency, f_c, for a signal at frequency f_i that could cause interference when the third harmonic ($3f_i$) mixes with the third harmonic of the local oscillator ($3f_{LO}$). There are two answers. Find them both and express your results in terms of f_{IF}.

18. (Image-reject mixer) When a multiplier is used to perform frequency conversion it is necessary to include a pre-selector filter in front of the multiplier to prevent signals at the image frequency from being converted to the intermediate frequency (IF) at the mixer output. The filter becomes hard to realize when the IF is small, since the image

frequency is very close to the desired frequency. When the IF is small, a better method for rejecting images is the phasing method, whereby the response from the image is cancelled at the mixer output. One implementation of a phasing-type image-reject mixer is shown in Figure 2.16.

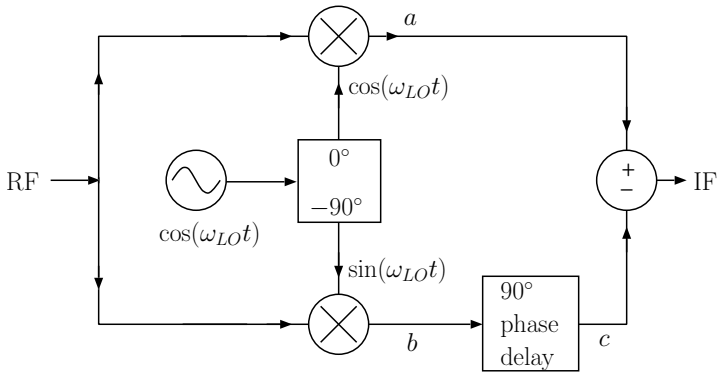


Figure 2.16: Image-reject mixer

(a) For the RF input signal $A\cos\omega_{RF}t$, find the time-domain IF output signal if the + sign is used in the signal combiner. You will need to consider two cases: (i) $\omega_{LO} > \omega_{RF}$ and (ii) $\omega_{LO} < \omega_{RF}$. Be careful - if you have a signal of the form $\sin[(\omega_1 - \omega_2)t]$ and you want to delay it by 90 degrees, then the result will be $\mp\cos[(\omega_1 - \omega_2)t]$ where the upper sign applies if $\omega_1 > \omega_2$ and the lower sign applies if $\omega_1 < \omega_2$.

(b) Repeat part a for the case where the − sign is used in the signal combiner.

(c) Use your results from parts (a) and (b) to answer this part. Suppose this mixer is to be used to convert a desired signal at frequency ω_c (i.e. RF input signal is $A\cos\omega_c t$) to an IF denoted by ω_{IF} using high LO ($\omega_{LO} = \omega_c + \omega_{IF}$). The image frequency is $\omega_{IM} = \omega_c + 2\omega_{IF}$. Which sign should be used in the combiner in order for the system to produce the desired frequency conversion?

(d) For the sign that you chose in part (c), write the IF output signal when the input signal's frequency is $\omega_{IM} = \omega_c + 2\omega_{IF}$, i.e. when the RF input signal is $A\cos\omega_{IM}t$. You should find that this input signal produces a finite output from the mixer. Why, then, is this called an image-reject mixer?

19. We have seen several systems that use cancellation to eliminate unwanted signal components. In practice, perfect cancellation is difficult to achieve because of inevitable amplitude and phase differences between the two signals that are being combined. In this problem, we analyze the impact of amplitude and phase imbalance on the suppression of unwanted terms. Consider the superposition of two signals which differ slightly in amplitude and phase, i.e.

$$s_{\pm}(t) = A\cos(\omega t) \pm (A + \delta A)\cos(\omega t + \delta\theta).$$

(a) Consider phase errors only ($\delta A = 0$) and find the maximum phase imbalance ($\delta\theta$), in degrees, if the difference signal is to be suppressed by at least 10, 30, and 60 dB relative to the sum signal. The difference signal suppression, in dB, is $S = 10\log(\frac{P_+}{P_-})$, where $P_\pm$ represents the time-average power in the sum and difference signals.

(b) Consider amplitude errors only and repeat part (a). The amplitude imbalance, expressed in dB, is $I = 20\log(\frac{A+\delta A}{A})$. Determine the maximum amplitude imbalance if the difference signal is to be suppressed by at least 10, 30, and 60 dB relative to the sum signal.

Chapter 3

Properties of Passive Components

3.1 High Frequency Characteristics of Components

Passive electronic components such as resistors, capacitors, and inductors all exhibit some amount of dielectric and/or ohmic loss, as well as energy storage in electric and magnetic fields within and surrounding the component. Accurate models for the impedance of real components must, therefore, include resistance, capacitance, and inductance. In addition, the *skin effect*, and inductance and capacitance associated with the conductors that connect the component to the rest of a circuit cannot be neglected. More generally, at sufficiently high frequencies, the concept of a *lumped element* must be replaced with that of a *distributed circuit*. An accurate equivalent circuit model that includes all of these effects may be significantly more complex than the low-frequency circuit model that includes only the basic circuit element. It is important to understand the nature of these effects, so that they can be accounted for when designing practical circuits. Some of the most useful equivalent circuit models for common components will be discussed in this section.

3.1.1 Wire Above a Ground Plane

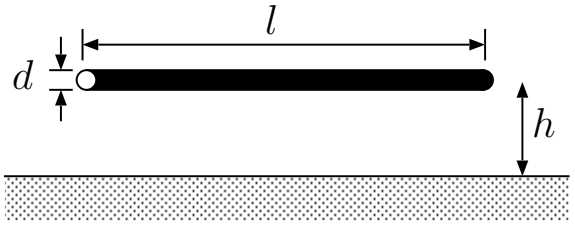


Figure 3.1: Wire above a ground plane.

Figure 3.1 shows a cylindrical wire parallel to a ground plane. A similar geometry may be used to model a trace on a circuit board if the cylindrical conductor is replaced with a flat conducting strip.

3.1.1.1 Resistance of Wires

Current tends to flow near the external surface of a conductor at high frequencies. The time-varying magnetic field associated with an initially uniform current density induces an electric field within the conductor which, in turn, drives a current against the initial current, forcing the net current density deep within the conductor to zero. As a result, the current density directed along the axis of a conducting wire is largest at the surface of the conductor, and falls to small values inside the conductor. Most of the current flows within the cylindrical shell within one skin depth from the surface. The skin depth is denoted by δ, where

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}},$$

f is frequency, in Hz, and μ, σ are the permeability and conductivity of the conductor. For non-magnetic conductors the permeability will be equal to the permeability of free space, $\mu = \mu_o = 4\pi \times 10^{-7}$H/m. Some values of conductivity are given in Table 3.1.

Material	σ S/m
Aluminum	3.5×10^7
Copper	5.8×10^7
Brass	1.5×10^7
Gold	4.1×10^7
Silver	6.1×10^7

Table 3.1: Conductivities for common metals.

The skin depth in copper at 100 kHz, 10 MHz, and 1 GHz is approximately 0.2 mm, 0.02 mm, and 0.002 mm, respectively. A typical wire used for connections and component leads is American Wire Gauge (AWG) 22 gauge wire, which has diameter $d = 25.3$ mils, or 0.643 mm. Hence, the skin depth in copper is a small fraction of the wire diameter at 10 MHz and higher frequencies.

At low frequencies, where the wire diameter is comparable to, or smaller than, the skin depth, then the DC resistance will be a good approximation to the wire resistance. The DC resistance is calculated using the total cross sectional area of the wire, $A = \pi d^2/4$, i.e.

$$R_{DC} = \frac{l}{\sigma A} = \frac{4l}{\pi d^2 \sigma}. \tag{3.1}$$

At higher frequencies, where the skin depth is small compared to the wire diameter, an accurate approximation for the resistance of the wire is obtained by assuming that the current flows with uniform density within one skin-depth, δ, from the wire surface, and with zero density elsewhere. The AC resistance is then obtained by replacing A in equation (3.1) with the area contained within one skin depth of the surface, i.e. $\pi d\delta$. The AC resistance can then be written as

$$R = R_{DC}\frac{\pi(d/2)^2}{\pi d\delta} = R_{DC}\frac{d}{4\delta}, \;\; \delta \ll d \tag{3.2}$$

where R_{DC} is the DC resistance of the wire. Since the skin depth decreases with increasing frequency, the resistance will increase as $f^{1/2}$. Obviously, the wire resistance can be much greater than the DC resistance, especially at high frequencies. For example, consider AWG

22 wire, which has DC resistance 0.0053 Ω/cm. At 10 MHz the resistance is ∼0.04 Ω/cm, and at 1 GHz the resistance is ∼0.4 Ω/cm.

The approximation given in equation (3.2) is based on the assumption that the wire is far from any other conductors, so that the time-varying magnetic field within the wire is only due to the current within the wire. Suppose that two parallel wires, carrying the same current, are brought into close proximity. In this case, the magnetic field within each wire will include a contribution from the current in the neighboring wire, i.e. the wires will be inductively coupled. In this case, the current density within the wire will not be uniformly distributed around the outer shell defined by the skin-depth. The electric field induced within the wires will cause the current density to be decreased on the side of the wire that is closest to the neighboring wire. This *proximity effect* reduces the effective cross-sectional area even further, and increases the AC resistance of the wire. The proximity effect can be very important in inductors where multiple wires are in close proximity.

3.1.1.2 Inductance of wires

The inductance, per unit length, of a wire with length l, diameter d, and distance h from a ground plane is (when $d \ll l$ and $h \ll l$):

$$L = \frac{\mu_0}{2\pi} \cosh^{-1} \frac{2h}{d}. \tag{3.3}$$

If the wire axis is located at least one wire diameter above the ground plane ($h/d > 1$), then the following approximation is useful

$$L \simeq \frac{\mu_0}{2\pi} \ln \frac{4h}{d}. \tag{3.4}$$

For values of h/d in the range 1 to 100, equation (3.4) predicts that L ranges from 2.8 nH/cm to 12 nH/cm. Notice that the inductance per unit length is relatively insensitive to the exact value of h/d.

A useful number to remember is that AWG 22 copper wire placed directly on the top surface of a printed circuit board with dielectric thickness of .062 inches =1.57 mm (a standard thickness), and with a ground plane on the bottom of the board, will have an inductance of approximately 4.8 nH/cm. At 100 MHz the inductive reactance of such a wire, $X_L = \omega L$, will be approximately 3 Ω/cm, and at 1 GHz the figure is 30 Ω/cm. For short wires, the resistance is often small enough to be ignored, however the inductance of the wire is often significant. In a circuit where impedances are relatively low, the series impedance of even a short connecting lead may have a significant impact on circuit performance. This leads to a fundamental rule of RF circuit design - at high frequencies it is important to keep the length of interconnecting wires and circuit-board traces short in order to minimize lead inductance. When components are separated by significant distances, interconnections must be treated as distributed circuit elements, and transmission line models are used to model the conductors that interconnect components.

3.1.2 Resistors

Several types of resistors are used in RF circuits, including wire-wound, carbon composition, thick film, and thin film units. Wire-wound resistors consist of a length of lossy wire that is coiled up to fit into a small package. This type of resistor is seldom used at RF because they

have relatively large inductance. Carbon composition resistors consist of a lossy dielectric material sandwiched between two conducting electrodes. Thick or thin-film resistors consist of a film of conducting material deposited on an insulating substrate. The film is in contact with two conductive electrodes to provide a means for connection to external circuitry. Film resistors are available in cylindrical packages with attached connecting leads and also as surface mount devices (SMDs). Thick or thin film resistors in a surface mount package (aka, "chip" resistors) are the most common type for RF applications.

Over a fairly wide frequency range, resistors can be modeled using the equivalent circuit shown in Figure 3.2.

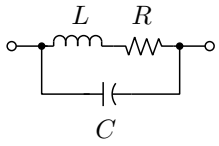


Figure 3.2: High frequency equivalent circuit for a resistor without external connecting leads.

The inductor in the model shown in Figure 3.2 represents the inductance associated with the current path through the resistor, and the capacitor represents the capacitance between the two electrodes used to connect the resistor to external circuitry. The unavoidable inductance and capacitance associated with the resistors (and other components) are sometimes termed *parasitic* inductance and capacitance. Both the inductance and capacitance of a resistor depend on the geometry and dimensions of the resistor. The inductance is primarily determined by the length of the current path within the element, and the capacitance is determined by the size and separation of the contact electrodes, as well as the dielectric permittivity of the material between the electrodes. Generally, the inductance and capacitance associated with a miniature surface mount resistor package are on the order of 1 nH and 1 pF, respectively. Significantly higher inductance would be associated with a part in an axial package with wire leads. Even with very short leads, such a package would have a typical inductance value on the order of 10 nH. If the component has external connecting leads attached to the package, then the inductance of the wire leads and the capacitance between the leads may need to be added to the model, as shown in Figure 3.3 .

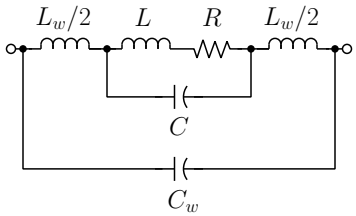


Figure 3.3: Model for a resistor with connecting leads. The total inductance of the leads is L_w and the capacitance between the leads is C_w.

For many purposes, the simpler model of Figure 3.2 can be used even for resistors with

external connecting leads if the inductance is taken to be the sum of the resistor inductance and the lead inductance, and the shunt capacitance is taken to be the sum of the package capacitance and the capacitance between the leads. The following discussion will be based on the simpler model shown in Figure 3.2.

When the resistance is small ($\ll 100\,\Omega$), and the frequency is not too high, the series LR branch of the model has a much lower impedance than the capacitance that shunts it. The capacitive reactance can then be neglected, leading to the simplified series RL model in Figure 3.4.

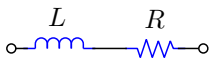


Figure 3.4: Equivalent circuit for a small resistance

For example, consider a 50 Ω resistor with lead inductance L = 10 nH, shunt capacitance C = 1 pF. Figure 3.5 shows the magnitude of the impedance versus frequency up through 1 GHz calculated using the full model. In this case the series inductance acts to increase the impedance at high frequencies.

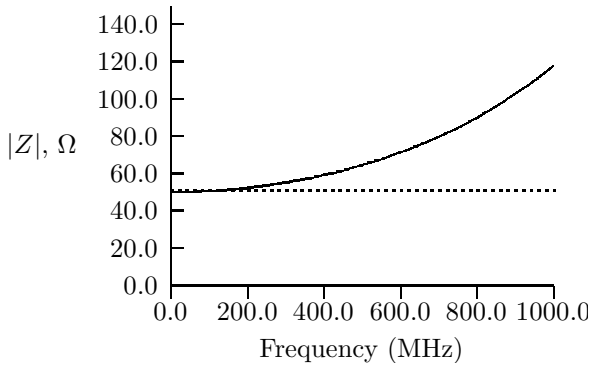


Figure 3.5: Impedance versus frequency of a 50 Ω resistor. The dotted line shows the impedance of an ideal resistor.

When the resistance is large ($\gg 100\,\Omega$) and the frequency is not too high the inductive reactance will be small compared to R and can be neglected; Figure 3.6 shows the equivalent circuit in this case.

For example, consider a 10 kΩ resistor with the same inductance and shunt capacitance as before (L=10 nH, C=1 pF). Figure 3.7 shows the impedance versus frequency calculated using the full model. Here the shunt capacitance is dominant and tends to "short out" the resistance, resulting in a dramatic reduction in the impedance at high frequencies.

For moderate resistance values the parasitic elements tend to have a smaller effect on the impedance. Figure 3.8 shows the magnitude of the impedance for a 200 Ω resistor with L = 10 nH, C = 1 pF. In this case neither of the parasitic reactances is significant compared to 200 Ω and the impedance variation is relatively small up through 1000 MHz.

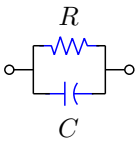


Figure 3.6: Equivalent circuit for a resistor with high resistance.

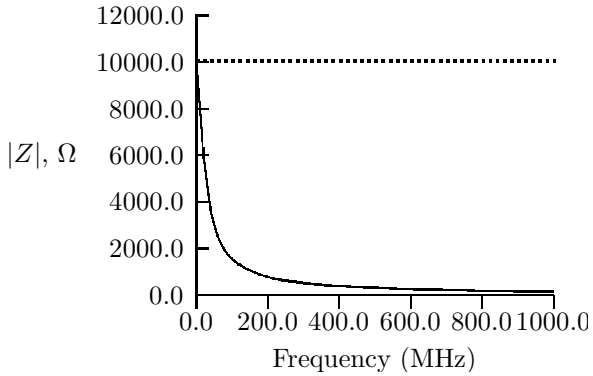


Figure 3.7: Impedance versus frequency of a 10 kΩ resistor. The dotted line shows the impedance for an ideal resistor.

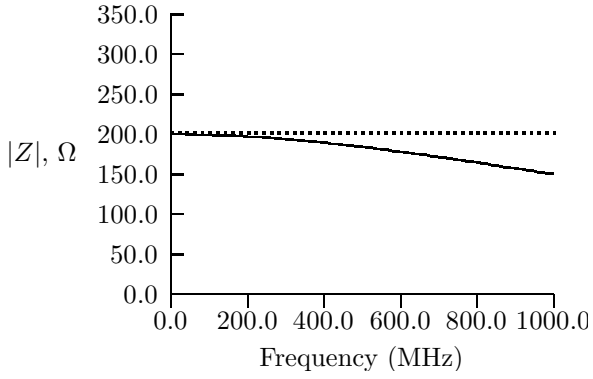


Figure 3.8: Impedance versus frequency of a 200 Ω resistor

Notice that the impedance of a resistor can be much larger or smaller than the DC resistance, depending on the resistance value and frequency. Until some intuition is developed, it is a good idea to measure the impedance of any resistor that is being considered for use in an RF circuit.

3.1.3 Capacitors

Capacitors are constructed by separating two conducting electrodes by an insulating medium such as air, or a low-loss dielectric material. Loss in the dielectric is modeled as a resistance (R_p) in parallel with the intrinsic capacitance. The inductance associated with the current path through the electrodes and any connected leads appears in series as shown in Figure 3.9. A resistance in series with the inductor (R_s) models losses in the electrodes and leads.

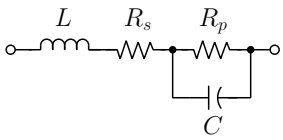


Figure 3.9: Capacitor model including lead inductance (L), dielectric loss resistance (R_p), and conductor loss (R_s).

The model can be transformed into a series RLC model using a parallel to series transformation, as shown in Figure 3.10. where it is assumed that the dielectric loss resistance is large compared to the capacitive reactance ($R_p \gg 1/(\omega C)$). The sum of the ohmic resistance and the transformed dielectric loss resistance is termed the equivalent series resistance (ESR), denoted here by R_{esr}, i.e.

$$R_{esr} = R_s + 1/(R_p \omega^2 C^2).$$

The ESR of a capacitor is dominated by dielectric loss at sufficiently low frequencies, whereas the resistance of the electrodes and leads will dominate at high frequencies.

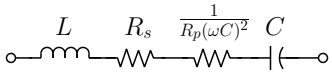


Figure 3.10: Equivalent series RLC circuit of a capacitor, valid where $R_p \gg 1/(\omega C)$, derived from the model shown in Figure 3.9.

The Q of a capacitor at any frequency is the ratio of the reactance and the ESR

$$Q = \frac{|X|}{R_{esr}}, \tag{3.5}$$

where the reactance $X = \omega L - 1/(\omega C)$. The *dissipation factor*, d, is the inverse of the Q, i.e

$$d = \frac{1}{Q} = \frac{R_{esr}}{|X|}. \tag{3.6}$$

The dissipation factor is also called the *loss tangent* because it is the complement of the tangent of the phase angle associated with the capacitor impedance.

$$\tan\delta = \frac{R_{esr}}{|X|} = \omega C R_{esr}$$

At frequencies well below the series resonant frequency of a capacitor, the reactance reduces to $X \simeq -1/(\omega C)$ and Q, d, and $\tan\delta$, can be written as follows:

$$Q = \frac{1}{\omega C R_{eq}}$$

$$d = \tan\delta = \omega C R_{eq}.$$

All capacitors will have a series resonant frequency, $f_s = 1/(2\pi\sqrt{LC})$. Above this frequency, the inductive reactance dominates, and the net reactance is positive. Hence, capacitors are inductive at frequencies above f_s! For a given package type the inductance will be roughly independent of the capacitance value, hence the series resonant frequency will be lower for larger values of capacitance.

As an example, consider a 0.01 μF capacitor with total lead length of 1 cm. We know, from previous discussion, that a typical value of the lead inductance is 10 nH/cm, which gives a total inductance of 10 nH and a series resonant frequency of 15.9 MHz The magnitude of the impedance and the reactance versus frequency are shown in Figure 3.11. The impedance goes to zero at the series resonant frequency in this lossless model. A real capacitor's impedance will fall to a minimum value equal to R_{esr} at the series resonant frequency, f_s.

Circuit designers make explicit use of the fact that capacitor impedance is smallest at and near the series resonant frequency. When a capacitor is used as a DC bias circuit decoupling element or as an interstage DC-blocking coupling element, it is sometimes possible to choose the capacitor so that the intended frequency of operation falls near the series resonant frequency of the capacitor, where the impedance is smallest.

3.1.4 Inductors

3.1.4.1 Air Core Inductors

An approximate formula for the inductance of a close-wound single-layer coil with nonmagnetic core (e.g., air) is

$$L = \frac{(rN)^2}{9r + 10l} \tag{3.7}$$

where

L = inductance in μH

N = number of turns

r = radius of coil (inches)

l = length of coil (inches)

Formula 3.7 is accurate to within 1% if $l > 0.8r$, i.e., if the coil is not too short.

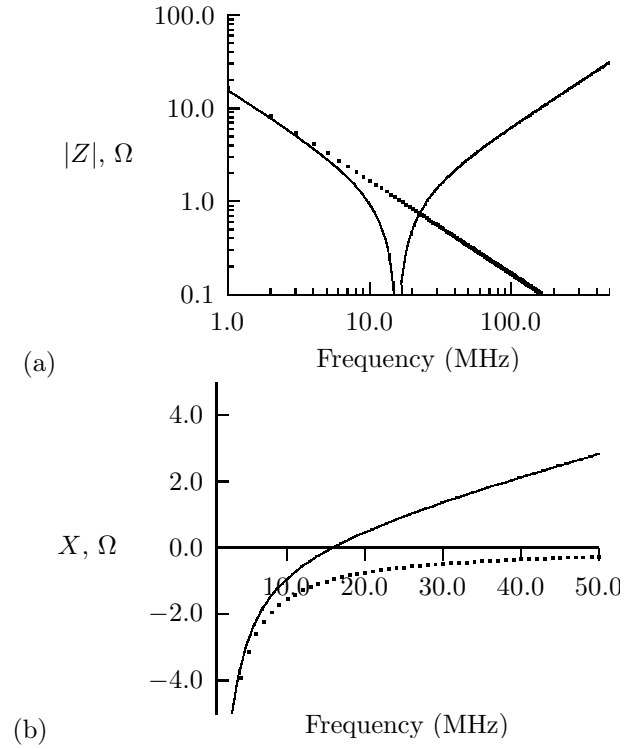


Figure 3.11: (a) Log-magnitude of impedance and (b) reactance for a 0.01 μF capacitor with series inductance of 10 nH. The dotted line shows the result for an ideal 0.01 μF capacitor.

3.1.4.2 Toroidal Inductors

Toroidal inductors are formed by winding a coil of wire around a donut-shaped (toroidal) core. An important advantage of this type of inductor is its self-shielding property. Typically, the core material will have high relative magnetic permeability ($\mu_r \gg 1$), and the magnetic field will be essentially confined to the core. This means that the coil inductance will be unaffected by its physical orientation and also that negligible mutual coupling will exist between toroidal coils in close proximity. Circuit design can be greatly simplified if mutual coupling effects can be neglected.

The core material for a toroidal inductor is usually a magnetic material, that is, a material with magnetic permeability larger than the permeability of free space. Inductor cores for RF applications are either manufactured from iron powder or ferrite material. The powdered iron cores consist of small iron particles suspended in an insulating compound. The mixture is compressed and baked at high temperature to produce a rigid structure. Different mixtures exhibit different magnetic permeabilities. Ferrites are ceramic materials with iron oxide as the dominant constituent. The iron oxide is mixed with nickel, manganese, zinc, or magnesium. Again, different mixtures exhibit different magnetic permeabilities. Generally, ferrite cores offer higher magnetic permeability than iron-powder cores, which results in fewer turns required to realize a given inductance value. Iron-powder cores generally have higher saturation flux densities. Saturation flux density refers to the largest flux density for which the linear relationship $B = \mu H$ holds. A general rule of thumb for high power RF applications is that the power handling capability of a ferrite core is limited by flux saturation, while the limiting factor for iron powder is temperature rise.

The important parameters of the core material are its relative magnetic permeability, cross-sectional area and diameter, and volume resistivity (which determines core loss). Other considerations for high power applications are the saturation flux density, which determines the largest magnetic field that the core can support, and the temperature rise resulting from power dissipation. For small signal applications these need not be considered, but the temperature stability of the core may need to be considered.

An approximate formula for the inductance of a toroidal winding having cross sectional area A and effective length l_{eff} is:

$$\begin{aligned} L &= \frac{\mu A}{l_{eff}} N^2 \\ &= A'_L N^2 \end{aligned} \tag{3.8}$$

where

$$\mu = \mu_r \mu_o \tag{3.9}$$

$$A'_L = \frac{\mu A}{l_{eff}} \tag{3.10}$$

The numerical value of the parameter A'_L is the inductance for a single-turn winding. The core manufacturer will provide the A'_L values for each type of core. The inductance of an N-turn winding is found from equation (3.8). Often, core manufacturers will characterize their cores by specifying the inductance of a winding with some particular number of turns.

For example, one manufacturer specifies the inductance per 100 turns for iron powder cores. If this constant is denoted by A_L, it is related to A'_L in equation (3.10) by

$$A_L = A'_L(100)^2. \tag{3.11}$$

On data sheets, the units of A_L would typically be given as as "μH/(100 turns)".

3.1.4.3 Equivalent circuit model for an inductor

A reasonably accurate equivalent circuit for an inductor consisting of a coil of wire wound as a solenoid, around a torus, or in a plane includes a series resistance to model the ohmic loss in the wire and a shunt capacitance to model the *distributed capacitance* between the turns of the coil. The inductance of the wire results in voltage differences between the different parts of the coil. This voltage difference sets up an electric field in the air and in any dielectric material near the coil. The effect of this stored electric energy can be modeled with a capacitance shunted across the terminals of the coil. This effective capacitance is called the distributed capacity of the coil. The distributed capacity of a coil is dependent on the geometry of the coil and the number of turns. In general, very closely spaced windings will have a larger distributed capacity. The equivalent circuit for an inductor is shown in Figure 3.12. If the inductor is wound on a core with finite conductivity (ferrite or iron-powder are often used), then it may be necessary to account for core losses by augmenting the model shown in Figure 3.12 with a resistor in parallel with the capacitor.

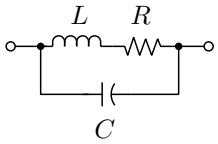


Figure 3.12: Equivalent circuit for an inductor

A lower-bound on the coil resistance may be estimated from the AC resistance of an isolated wire having length equal to the total length of wire in the coil, i.e.

$$R = r_{DC}\frac{n\pi dD}{4\delta}\ \Omega \tag{3.12}$$

where r_{DC} is the DC resistance per unit length of the wire, d is the wire diameter, D is the coil diameter, n is the number of turns in the coil, and δ is the skin depth in the wire. A more accurate estimate of the coil resistance can be obtained by accounting for the proximity effect. The importance of the proximity effect will depend on the geometry of the coil. The proximity effect may cause the actual resistance to be significantly larger than the estimate given in equation 3.12.

The impedance and reactance versus frequency are shown in Figures 3.13 and 3.14 for a 10 μH inductor with (constant) series resistance of 15 Ω and distributed capacity of 20 pF. Note the parallel resonant frequency $f_p \simeq 1/(2\pi\sqrt{LC})$ at which the impedance of the inductor has a relatively large value.

Circuit designers often make use of the fact that an inductor has a very large impedance near the parallel resonant frequency, and will deliberately use the device at or just below

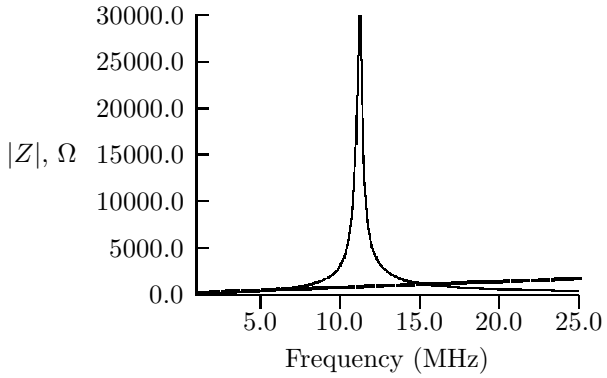


Figure 3.13: Impedance versus frequency for a 10 μH inductor with series resistance of 15 Ω and distributed capacity of 20 pF. The dotted line shows the impedance for an ideal 10 μH inductor.

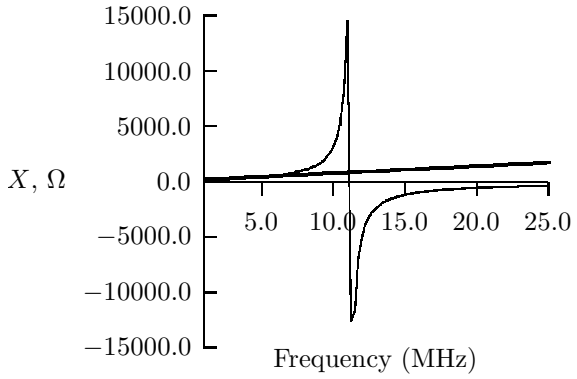


Figure 3.14: Reactance versus frequency for a 10 μH inductor with series resistance of 15 Ω and distributed capacity of 20 pF. The dotted line shows the reactance for an ideal 10 μH inductor.

this frequency. For example, in many cases an inductor is used to provide a DC bias signal to a circuit, but it is necessary to isolate the bias supply from the circuit at the operating frequency. In this case, an inductor with parallel resonant frequency near the operating frequency may be employed to provide a low impedance to DC and large impedance to RF signals. In such an application the inductor element is referred to as an *RF choke*. It should be noted that the parallel resonant frequency of an inductor depends critically on its construction. It is difficult to accurately estimate the parallel resonant frequency of an inductor. In practice it is usually necessary to measure the parallel resonant frequency as well as the series resistance.

When considering a particular inductor for use in a circuit, the designer needs to be aware of the parallel resonant frequency as well as the "Quality Factor," or Q, of the inductor. The Q of an inductor is defined to be the ratio of inductive reactance and resistance associated with the component, i.e.,

$$Q = \frac{|X_s|}{R_s} \tag{3.13}$$

where the impedance of the inductor is $Z = R_s + jX_s$. The higher the Q, the better the inductor approximates an "ideal" component. The Q is an important parameter if the inductor is to be used in a resonant circuit, filter, or matching network.

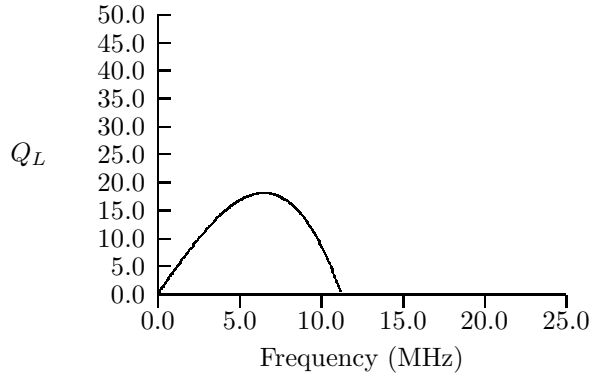


Figure 3.15: The quality factor ($Q_L = |X_S|/R_S$) for the inductor whose impedance characteristics are plotted in Figures 3.13 and 3.14. The component is inductive only at frequencies below its self-resonant frequency (approx. 11.2 MHz in this case), so the Q is not plotted at frequencies above the self-resonant frequency. The optimum frequency of operation for an inductor is at or near the peak of the Q vs. frequency curve.

Finally, it should be noted that substantial mutual inductance may exist between inductors that are located in close proximity. This coupling can cause serious problems if it has not been accounted for in the design of the system, especially when it occurs between the input and output circuits of an amplifier, since unwanted oscillation may result. Mutual coupling effects can be minimized by using self-shielding construction (e.g., toroidal inductors) or, in the case of air-wound coils, by orienting the axes of the coils so that they are perpendicular. A model and theoretical analysis for coupled inductors will be presented in a later chapter.

3.2 References

1. Bowick, Chris and Howard W. Sams, *RF Circuit Design*, Indianapolis, Indiana, 1982.

2. DeMaw, M. F., *Ferromagnetic Core Design and Application Handbook,* Prentice Hall, Inc, 1981.

3. Ludwig, Reinhold and Pavel Bretchko, *RF Circuit Design - Theory and Applications*, Prentice Hall, 2000.

4. Terman, Frederick Emmons, *Radio Engineers Handbook*, McGraw Hill, 1943.

3.3 Homework Problems

1. You are given a black box with two terminals. Suppose that you know that the box contains a passive circuit that is constructed from 3 elements: a resistor (R), lossless capacitor (C), and lossless inductor (L). Your task is to figure out how the elements are connected, and what their values are. You make some measurements of the impedance of the box $Z(f) = R(f) + jX(f)$. The results of the measurements are shown in Figure 3.16. Sketch the circuit that is inside of the box. Estimate the

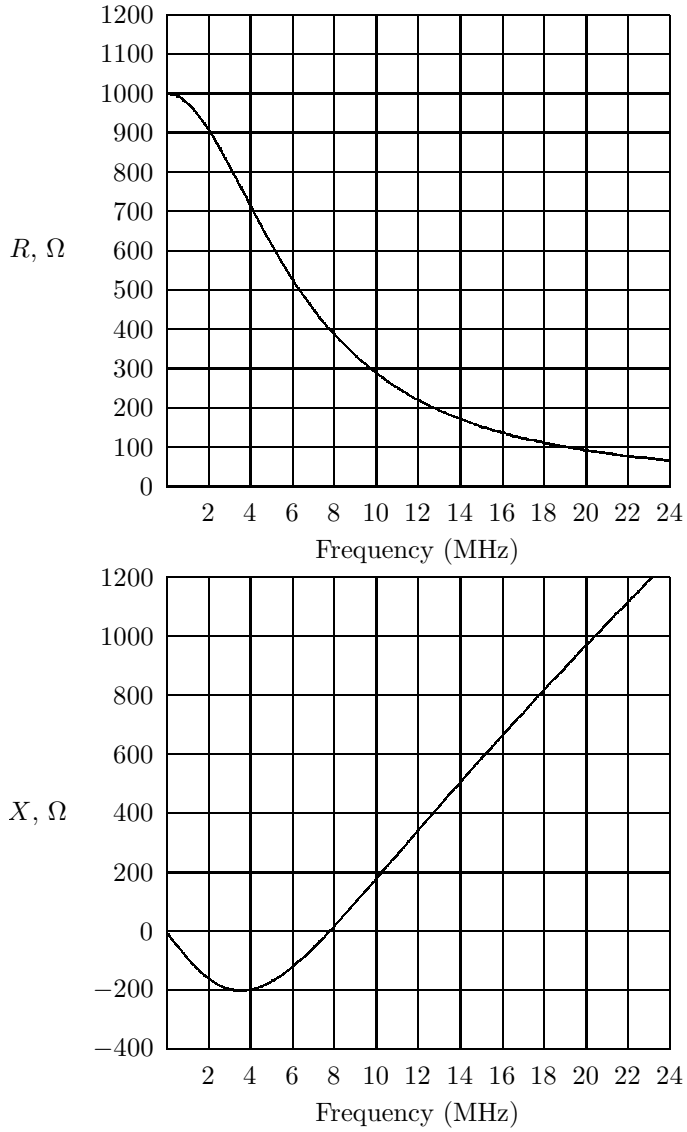


Figure 3.16: Measured resistance (top) and reactance (bottom).

element values. You may assume that the elements (R, L, C) are ideal, i.e., assume that distributed capacitance across the inductor and the lead inductance of the resistor and capacitor can be neglected.

2. Same as problem 1, except the measured data is shown in Figure 3.17. Sketch the

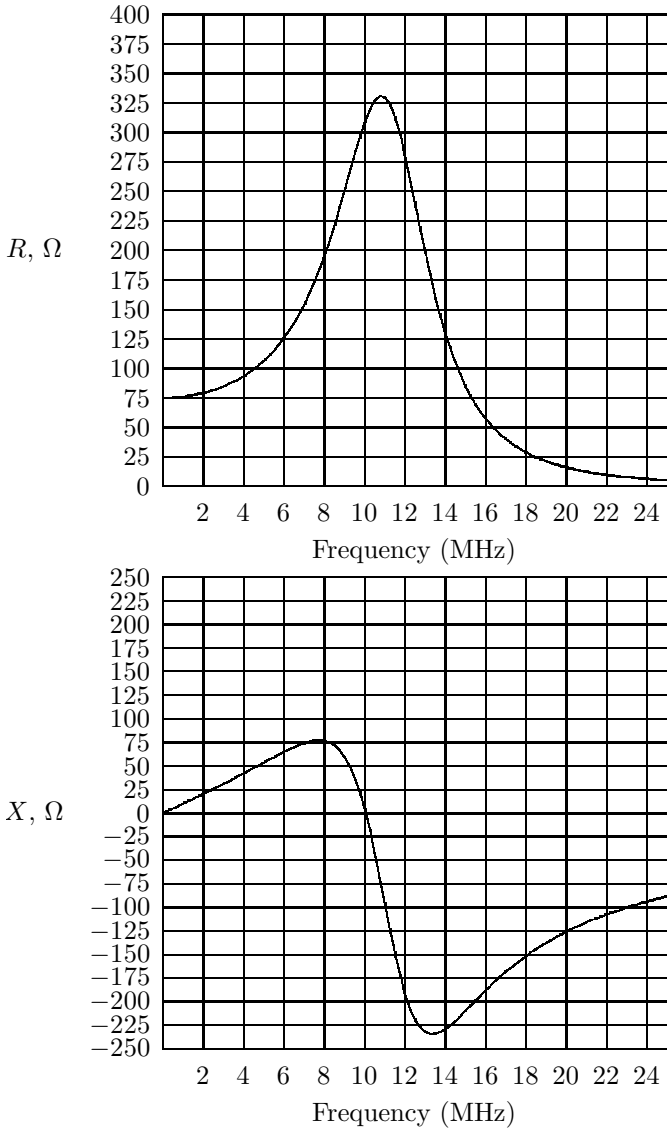


Figure 3.17: Measured resistance (top) and reactance (bottom).

circuit that is inside of the box. Estimate the element values. You may assume that the elements (R, L, C) are ideal, i.e., assume that distributed capacitance across the inductor and the lead inductance of the resistor and capacitor can be neglected.

3. Same as problem 1, except the measured data is shown in Figure 3.18. Sketch the

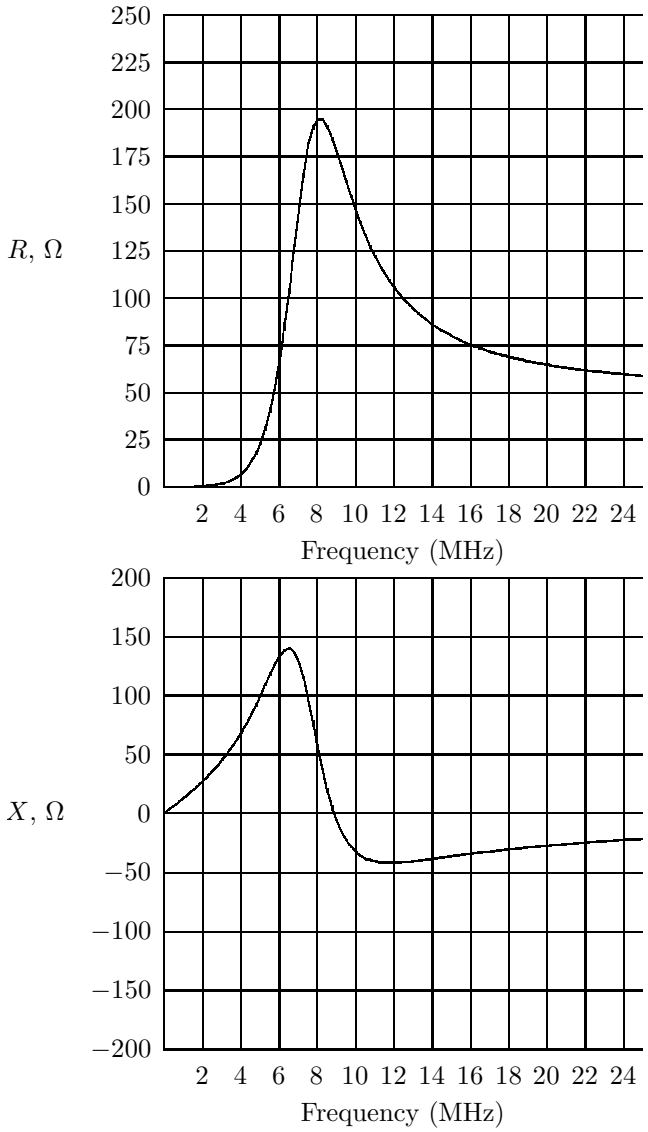


Figure 3.18: Measured resistance (top) and reactance (bottom).

circuit that is inside of the box. Estimate the element values. You may assume that the elements (R, L, C) are ideal, i.e., assume that distributed capacitance across the inductor and the lead inductance of the resistor and capacitor can be neglected.

4. The circuit shown in Figure 3.19 is used as a model for a realistic resistor or inductor. It can also be used to model a realistic parallel resonant circuit.

   (a) Find an expression for the impedance of this circuit.

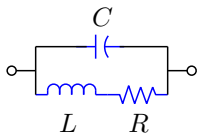


Figure 3.19: Model for realistic inductor and for parallel resonant circuit

(b) The frequency at which the impedance is purely resistive is called the "resonant frequency." Under what conditions will there be a frequency (> 0) at which the impedance is purely resistive? Show that if this condition is satisfied, and if $CR^2/L \ll 1$, then the resonant frequency is given approximately by:

$$\omega_o = \frac{1}{\sqrt{LC}}(1 - \frac{CR^2}{2L}) \tag{3.14}$$

Note that this frequency is the parallel resonant frequency of the circuit and that the effect of the resistance, R, is to make the parallel resonant frequency slightly smaller than that of the lossless (R=0) case.

(c) Continuing from part 4b, assume that the term CR^2/L can be ignored and find the magnitude of the impedance at the parallel resonant frequency. How does it depend on R?

(d) For parts 4d and 4e you should start with the exact expression for the impedance. Find an approximate expression for the impedance valid when $\omega \ll \frac{1}{\sqrt{LC}}$ and $\frac{L}{R^2C} \gg 1$. Draw a simplified equivalent circuit that is valid under these conditions.

(e) Find an approximate expression for the impedance valid when $\omega \ll \frac{1}{\sqrt{LC}}$ and $\frac{L}{R^2C} \ll 1$. Draw a simplified equivalent circuit that is valid under these conditions.

5. Define the "effective capacitance", C_{eff}, of a realistic capacitor having finite lead inductance to be the capacitance of the ideal capacitor that has the same reactance as the real capacitor. C_{eff} will depend on frequency.

   (a) Find an expression for the effective capacitance of a realistic capacitor. Express your result in terms of the capacitance, C, and the resonant frequency, f_o of the real capacitor.

   (b) What is the effective capacitance of a 300 pF capacitor with 20 nH of lead inductance at 10 MHz and at 40 MHz?

6. Figure 3.12 shows the equivalent circuit for an inductor.

   (a) Derive an exact expression for the self-resonant frequency of the inductor. The self resonant frequency is defined to be the frequency (> 0) where the impedance of the inductor is purely resistive. Express your result in terms of R, L, and C. When is $\frac{1}{2\pi\sqrt{LC}}$ a good approximation to f_p?

   (b) Consider a 500 nH inductor with series resistance $R = 1\ \Omega$ and shunt capacitance $C = 1$ pF. Find the self-resonant frequency of the inductor, f_p. Express your result in MHz.

(c) For the inductor specified in part (b), find the impedance of the inductor at f_p.

(d) For the inductor specified in part (b), find the "effective inductance", L_{eff}, of the inductor at 0.75 f_p. L_{eff} is defined to be $X_L(\omega)/\omega$, where $X_L(\omega)$ is the reactance of the inductor at frequency ω. Express your result in nH.

Chapter 4

RLC Networks, Resonance, and Q

4.1 Series RLC Network

Consider the series RLC circuit in a filter configuration where the output voltage is taken across the resistor, as shown in Figure 4.1. The voltage transfer function is

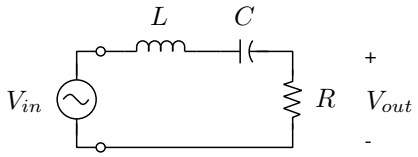


Figure 4.1: Series RLC circuit as a filter.

$$H(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{R}{R + sL + \frac{1}{sC}} \tag{4.1}$$

We will consider sinusoidal excitation under steady-state conditions, in which case we are interested in the frequency response, $H(j\omega)$:

$$H(j\omega) = \frac{R}{R + j\omega L\left(1 - \frac{1}{\omega^2 LC}\right)} \tag{4.2}$$

When $\omega = 1/\sqrt{LC}$, the phase shift of the transfer function is zero; this is called the "resonant frequency," ω_o, of the network and is the frequency at which the inductive and capacitive reactances are exactly equal in magnitude and, consequently, cancel each other:

$$\omega_o = \frac{1}{\sqrt{LC}} \tag{4.3}$$

The transfer function depends on R, L, and C, but only two parameters are necessary to specify the characteristics of the function. Define another quantity Q_s where:

$$Q_s = \frac{\omega_o L}{R} = \frac{1}{R}\sqrt{\frac{L}{C}} \tag{4.4}$$

The frequency response function can be rewritten in terms of only ω_o and Q_s:

$$H(j\omega) = \frac{1}{1 + jQ_s\left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}\right)} \tag{4.5}$$

The parameter Q_s is referred to as the series resonant circuit "Q." In a subsequent section it will be shown that the inverse of this quantity tells us what fraction of the total energy stored in the RLC circuit is dissipated in one complete cycle of the resonant frequency.

The magnitude and phase of the voltage transfer function (Equation 4.5) are plotted as a function of ω/ω_o in Figure 4.2(a) and (b) for Q=2 and Q=10. The same data is plotted with a logarithmic frequency axis in Figure 4.3(a) and (b).

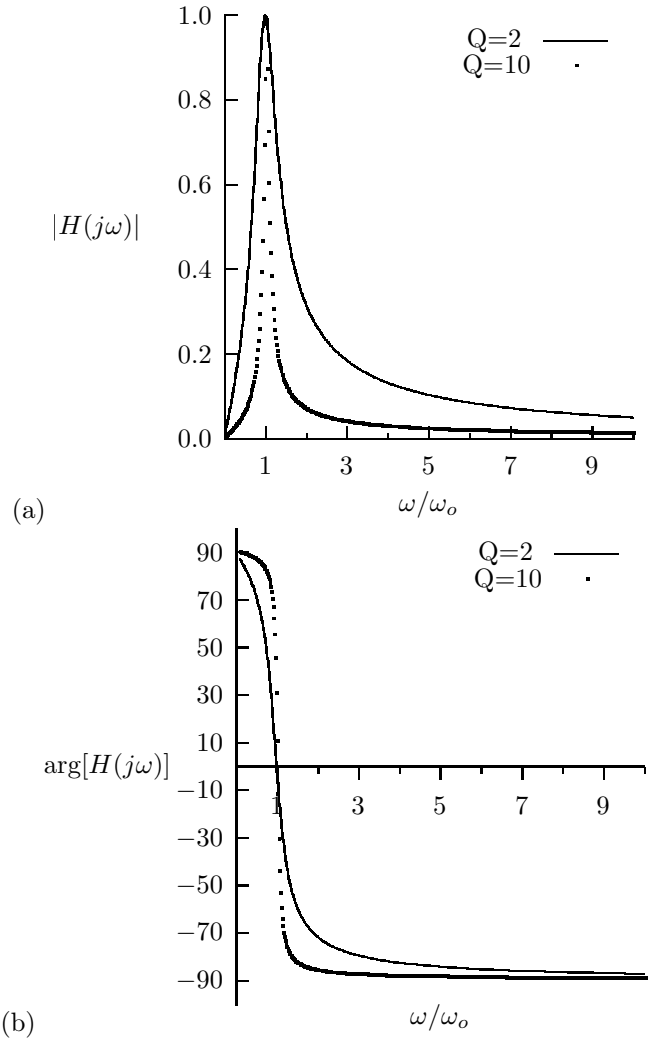


Figure 4.2: (a) Magnitude and (b) phase of the voltage frequency response for Q=2 (solid line) and Q=10 (dotted line).

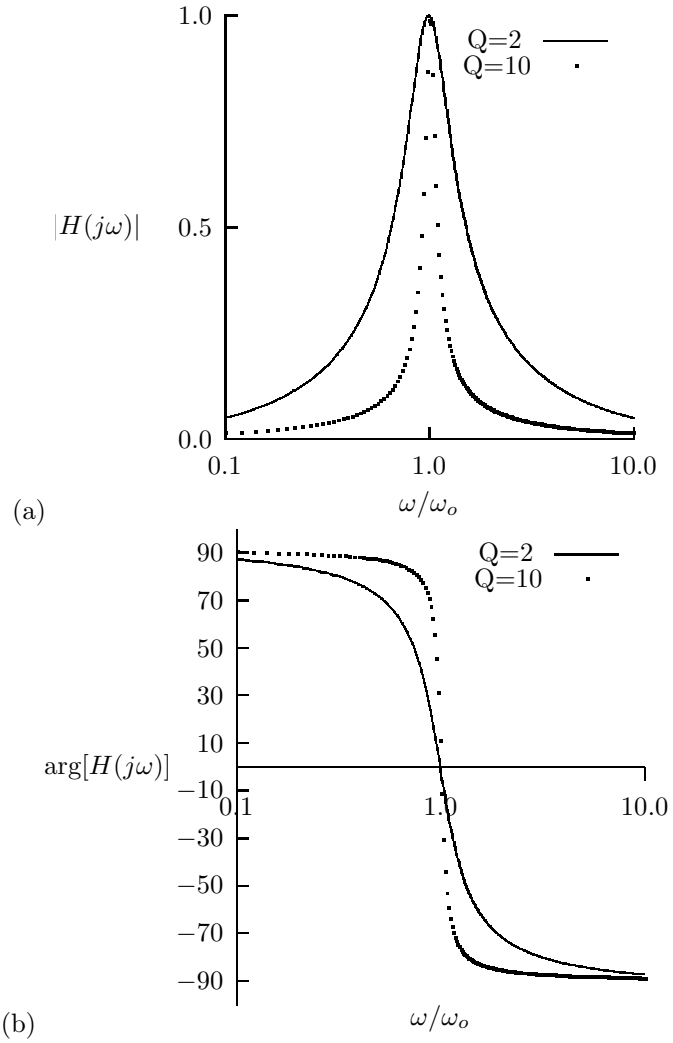


Figure 4.3: (a) Magnitude and (b) phase of the voltage frequency response for Q=2 (solid) and Q=10 (dotted) vs. ω/ω_o with a logarithmic frequency axis.

The RLC filter has a "bandpass" characteristic. The separation between the half-power (-3 dB) frequencies is often used to specify the bandwidth of a filter. The 3 dB bandwidth of the filter can be found by determining the difference between the frequencies where $|H(j\omega)| = 0.707$. Denote the lower and upper -3 dB frequencies by ω_1 and ω_2 as shown in Figure 4.4. Then $\Delta\omega = (\omega_2 - \omega_1)$ is referred to as the "3 dB bandwidth" of the filter. It is left as an exercise to show that

$$\Delta\omega = (\omega_2 - \omega_1) = \frac{\omega_o}{Q_s} \tag{4.6}$$

The bandwidth of a series RLC filter is inversely proportional to the Q_s of the circuit.

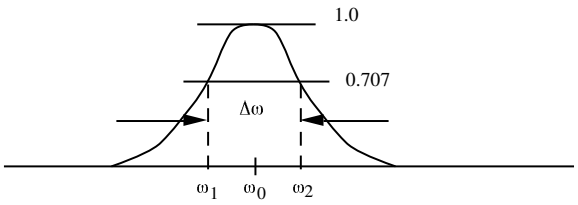


Figure 4.4: 3dB bandwidth of filter

4.1.1 Example - Series RLC circuit as a filter.

Use a series RLC circuit to couple a voltage source with negligible source resistance to a 50 Ω load as shown in Figure 4.5. The circuit should have a center frequency of 5 MHz and a 3 dB bandwidth of 100 kHz. The bandwidth and center frequency determine Q_s :

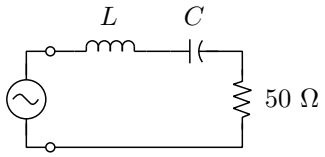


Figure 4.5: Circuit with 5 MHz center frequency and a 3 dB bandwidth of 100 kHz

$$Q_s = \frac{f_o}{\Delta f} = \frac{5\,\text{MHz}}{100\,\text{kHz}} = 50 = \frac{\omega_o L}{R} = \frac{2\pi(5\times10^6)L}{50} \tag{4.7}$$

so

$$L = \frac{50\,50}{2\pi(5\times10^6)} = 79.6\,\mu\text{H}$$

$$C = \frac{1}{\omega_o^2 L} = 12.7\,\text{pF}$$

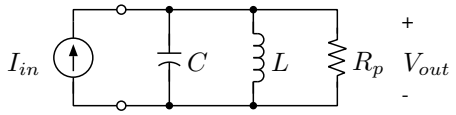


Figure 4.6: Parallel RLC as a filter

4.2 Parallel RLC

A parallel RLC circuit being driven by an ideal current source is shown in Figure 4.6. In this application the input current and output voltage are related by an impedance function, i.e.,

$$\frac{V_{out}(s)}{I_{in}(s)} = Z(s) \quad \text{(impedance)} \tag{4.8}$$

$$Z(s) = \left[\frac{1}{R_p} + \frac{1}{sL} + sC\right]^{-1}$$

For sinusoidal steady-state excitation

$$Z(j\omega) = \frac{R_p}{1 + jQ_p\left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}\right)} \tag{4.9}$$

where

$$\omega_o = \frac{1}{\sqrt{LC}} \tag{4.10}$$

$$Q_p = \frac{R_p}{\omega_0 L} \tag{4.11}$$

$$= \sqrt{\frac{C}{L}} R_p$$

This transfer function has exactly the same form as that of the series RLC circuit except for the scaling factor, R_p. Note, however, that the "Q" is defined differently for the parallel RLC. As before, the 3dB bandwidth is

$$\Delta\omega = \frac{\omega_o}{Q_p} \tag{4.12}$$

4.2.1 Unloaded vs Loaded Q of RLC circuits

If the source has a non-negligible impedance as shown in Figure 4.7 then

$$\frac{V_{out}}{I_{in}} = \frac{R_p \| R_S}{1 + jQ_p'(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})} \tag{4.13}$$

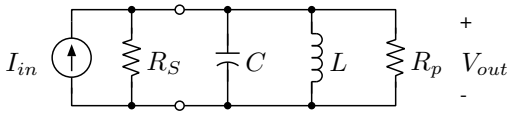


Figure 4.7: Parallel RLC filter driven by a source with finite source impedance.

where

$$Q_p^{'} = \frac{R_p \| R_S}{\omega_o L} \tag{4.14}$$

Compared to the case with infinite source impedance, the finite source impedance causes the Q to be reduced and, hence, the bandwidth to be increased. It is common practice to call the Q of the resonant circuit alone (either series or parallel RLC) the *unloaded* Q, and the Q of the composite circuit, which includes the source resistance and any other resistances that are external to the LC resonator, the *loaded* Q. The loaded Q is always smaller than the unloaded Q.

In the case of a series resonant circuit with resistance R driven by a source with finite impedance R_S, the loaded Q becomes

$$Q_s' = \frac{\omega_o L}{R_S + R}.$$

Again, the loaded Q is smaller than the unloaded Q.

4.3 More on Q

The previously defined Q's (Q_s, Q_p) describe the frequency selectivity of the simplest type of *resonant* RLC networks. The term "Q" has a more general interpretation which can be applied to any type of system that contains energy storage elements and dissipation. The general definition of Q for a system is

$$\begin{aligned} Q &= 2\pi \frac{\text{Maximum instantaneous stored energy}}{\text{Energy dissipated per cycle}} \\ &= 2\pi f \frac{\text{Maximum instantaneous stored energy}}{\text{Time} - \text{average power dissipated}} \end{aligned} \tag{4.15}$$

This definition can be applied to resonant and nonresonant circuits. If this energy-based definition is applied to resonant second-order RLC circuits the result is compatible with the (Q_s, Q_p) defined in the previous section. In principle, the definition could be applied to higher-order circuits, e.g, circuits with more than one inductor and capacitor, however this is usually not very useful. On the other hand, higher order circuits are often constructed by combining second-order circuits. In such cases, it is useful to characterize each of the constituent second-order resonant circuits by a "Q". This is common practice in filter-design, where multiple LC resonators are coupled to form a more complex filter. The energy definition is also commonly applied to characterize lossy inductors or capacitors which, by themselves, are non-resonant.

We shall first show that the energy definition is consistent with the resonant-circuit Q_s that has already been defined for series RLC circuits. Consider the series RLC with sinusoidal excitation as shown in Figure 4.8.

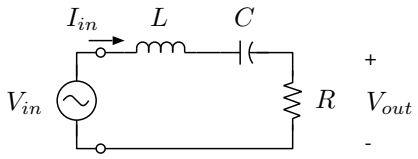


Figure 4.8: Series RLC circuit for calculation of Q

At resonance,

$$V_{out} = V_{in} \tag{4.16}$$

$$I_{in} = V_{in}/R \tag{4.17}$$

The fact that the voltage across the capacitor is 90 degrees out of phase with the current through the inductor means that the current maximizes at the time when the capacitor voltage is zero. At that instant in time, all of the stored energy resides in the inductor and the magnitude of the current phasor (which is the peak current magnitude) can be used to calculate the total stored energy. The stored energy is:

$$E_{max} = \frac{1}{2}L\left|I_{in}\right|^2 = \frac{1}{2}L\frac{\left|V_{in}\right|^2}{R^2} \tag{4.18}$$

Alternatively, at the time instant when the capacitor voltage is maximum then the current in the system is zero and all of the stored energy resides in the capacitor. The magnitude of the capacitor voltage phasor can then be used to calculate the stored energy at that time:

$$E_{max} = \frac{1}{2}C|V_{cap}|^2 = \frac{1}{2}C|\frac{I_{in}}{\omega_o C}|^2 = \frac{1}{2}L|I_{in}|^2 \tag{4.19}$$

The stored energy comes out the same either way. Actually, it can be shown that the total stored energy in this driven resonant RLC circuit is a constant, so that the maximum instantaneous stored energy is equal to the energy stored at any instant of time. The time-averaged power delivered to (and dissipated in) the network is:

$$\begin{aligned} P_{avg} &= \frac{1}{2}\mathrm{Re}[V_{in}I_{in}^*] \\ &= \frac{1}{2}\left|V_{in}\right|^2/R \end{aligned} \tag{4.20}$$

Then using the energy definition of Q:

$$Q_s = 2\pi f_o\,\frac{E_{max}}{P_{avg}} = \frac{\omega_o L}{R} \tag{4.21}$$

This result is the same as the definition for the series resonant Q that was given earlier.

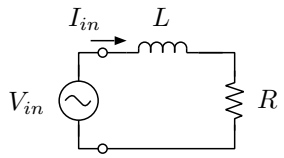


Figure 4.9: Nonresonant RL circuit.

It is also common to apply the definition to a nonresonant circuit, e.g., consider an RL circuit as in Figure 4.9.

$$I_{in} = \frac{V_{in}}{R + j\omega L} \tag{4.22}$$

In this case, the inductor is the only energy storage element. Because the current through the inductor varies sinusoidally, the stored energy will oscillate between zero and the maximum value given by:

$$\begin{aligned} E_{max} &= \frac{1}{2} L \left| I_{in} \right|^2 \\ &= \frac{1}{2} L \left| V_{in} \right|^2 \frac{1}{r_s^2 + \omega^2 L^2} \end{aligned} \tag{4.23}$$

The time-averaged power delivered is

$$\begin{aligned} P_{avg} &= \frac{1}{2} \mathrm{Re}[V_{in} I_{in}^*] \\ &= \frac{1}{2} \left| V_{in} \right|^2 \mathrm{Re}\left[\frac{1}{r_s - j\omega L} \right] \\ &= \frac{1}{2} \left| V_{in} \right|^2 \frac{r_s}{r_s^2 + \omega^2 L^2} \end{aligned} \tag{4.24}$$

Therefore

$$Q_L = 2\pi f \frac{E_{max}}{P_{avg}} = \frac{\omega L}{r_s} \tag{4.25}$$

In this case ω can be any frequency. This type of Q will be called the *component Q*. In this example, the RL circuit could represent a model for a lossy inductor. The component Q of an inductor evaluated at some frequency ω can be thought of as the *resonant Q_s or Q_p that results if a lossless capacitor is added to form a resonant circuit at the frequency ω.*

The concept of component Q is often used to describe the properties of arbitrary circuit elements at a particular frequency. For example, if an arbitrary circuit element can be represented by a series impedance $Z = R_s + jX_s$ at some frequency, then applying the definition of Q to that component yields $Q = |X_s|/R_s$ as illustrated in Figure 4.10(a). For a parallel representation of a circuit branch the component Q is as shown in Figure 4.10(b).

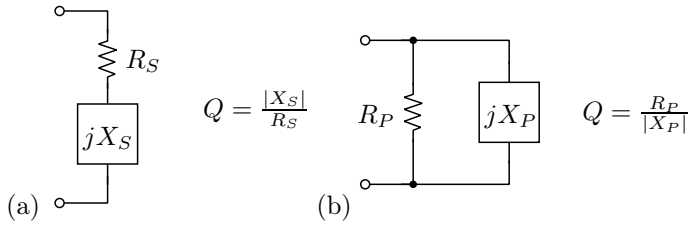


Figure 4.10: Definition of Q applied to a branch represented in (a) series form and (b) parallel form.

4.4 Series-to-Parallel Transformations

Any circuit element has both a series and a parallel representation. Since the energy storage and dissipation properties of the element do not depend on how we represent it, the Q is the same for either representation. The component Q concept is useful for series-to-parallel impedance transformations. For example, suppose the series impedance representation for a circuit element is known at a particular frequency as shown in Figure 4.11(a).

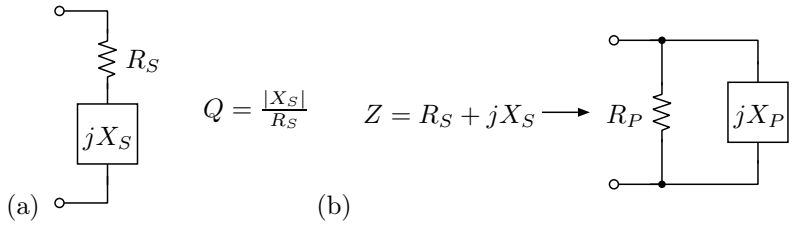


Figure 4.11: (a) Series impedance representation for a circuit at a particular frequency and (b) equivalent parallel representation having the same impedance as the original series branch.

The circuit Q is given by

$$Q = \frac{|X_s|}{R_s} \tag{4.26}$$

The equivalent parallel representation for the circuit element is shown in Figure 4.11(b). The equivalent parallel resistance and reactance are easily found after equating the impedances of the two models:

$$\begin{aligned} R_p &= R_s(1 + Q^2) \\ X_p &= X_s\left(1 + \frac{1}{Q^2}\right) \end{aligned} \tag{4.27}$$

Notice that the Q of the equivalent parallel circuit is $R_p/|X_p|$ which is equal to Q. Thus, the Q of the equivalent parallel representation is the same as that of the original series

representation (and vice versa). A useful simplification results if $Q \gg 1$, in which case

$$R_p \simeq R_s\, Q^2 \tag{4.28}$$

$$X_p \simeq X_s$$

Clearly, a complex impedance that is represented in parallel form, i.e. as $Z_p = R_p || jX_p$ can be transformed to series form, $Z_s = R_s + jX_s$ by defining the Q of the parallel representation ($Q = R_p/|X_p|$) and then calculating R_s and X_s from

$$R_s = \frac{R_p}{1+Q^2} \tag{4.29}$$

$$X_s = \frac{X_p}{1+\frac{1}{Q^2}} \tag{4.30}$$

4.4.1 Example - Series to parallel conversion

A 1 μH inductor has a component Q of 100 at 10 MHz. Find a parallel representation for the inductor. At 10 MHz, $\omega L = 2\pi(10^7)\,(10^{-6}) = 62.8\,\Omega$. So $Q_L = 100 = \frac{\omega L}{r_s} = \frac{62.8}{r_s}$.

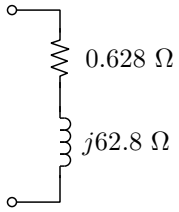


Figure 4.12: Representation of 1 μH inductor with a component Q of 100 at 10 MHz

Therefore, $r_s = 0.628\,\Omega$. Since $Q_L \gg 1$,

$$X_p \simeq X_s = 62.8\,\Omega \tag{4.31}$$

$$R_p \simeq Q^2\, r_s = 10^4(.628) = 6.28\,k\,\Omega$$

The equivalent network at 10 MHz is shown in Figure 4.13. It is important to remember that this equivalent circuit is only valid at 10 MHz.

The following example illustrates how a reactance in series with a resistor can form the basis of an impedance-transforming network that transforms the resistance into a different value.

4.4.2 Example - Impedance transformation

Design a lossless network to transform 50 Ω to 300 Ω using a capacitor and an inductor. We can start by putting an inductive reactance in series with the 50 Ω resistor, as shown on the left in Figure 4.14. The equivalent parallel representation, on the right in Figure

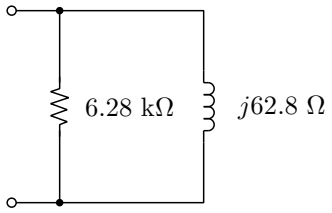


Figure 4.13: Equivalent network at 10 MHz.

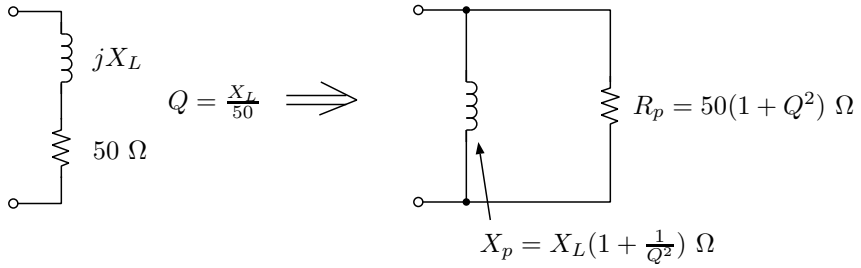


Figure 4.14: Inductive reactance in series with a 50 Ω resistor.

4.14, shows that the parallel resistance is larger than 50 Ω by the factor $1 + Q^2$. Setting $R_p = 300 = 50(1 + Q^2)$ yields the Q of the circuit ($Q = 2.24$) and therefore the value of the series reactance, $X_L = 2.24(50) = 112\,\Omega$. So far, it is known that the series circuit shown on the left in Figure 4.15 is equivalent to the parallel circuit shown on the right. Now a shunt

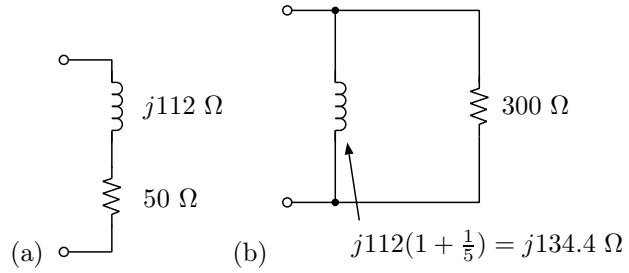


Figure 4.15: (a) Series circuit and (b) equivalent parallel circuit.

capacitor can be added to cancel the inductive reactance, as shown in Figure 4.16. Going back to the original, series, representation yields the final solution for the lossless network that transforms a 50 Ω resistance into a 300 Ω resistance as in Figure 4.17. This is an example of an L-section matching network which will be discussed in more detail in chapter 6. It provides an impedance match between 50 Ω and 300 Ω only at the design frequency. The circuit topology of this matching network is of the "lowpass" type; frequencies much

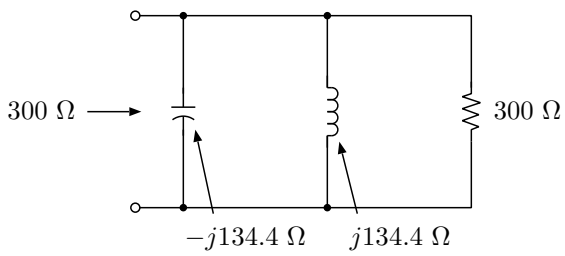


Figure 4.16: Shunt capacitor added to cancel inductive reactance.

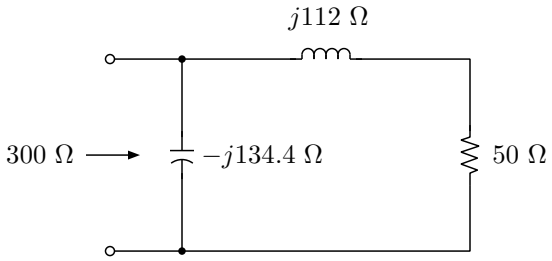


Figure 4.17: Lossless network to transform a 50 Ω resistance into a 300 Ω resistance.

lower than the design frequency will be transferred to the load, while frequencies much higher than the design frequency will not. Another solution to this matching problem exists and has a "highpass" topology as seen in Figure 4.18. Recall that the series resistance is

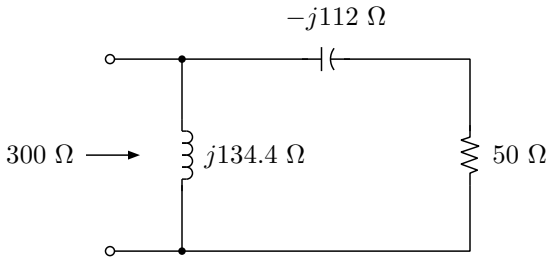


Figure 4.18: Highpass circuit topology.

stepped up to a larger parallel resistance by the factor $1 + Q^2$, which is independent of the sign of the series reactance. Thus, the highpass solution could be derived by starting by adding a capacitive reactance in series with the 50 Ω resistor.

4.4.3 Single-resonator filters

The impedance-transforming properties of a reactance in parallel or series with a resistance can be exploited to design simple RLC filters for a desired bandwidth when the source and load resistances and inductor (or capacitor) value can not be adjusted to produce the desired $Q_{s,p}$. For example, suppose that the source and load resistances are each $50\,\Omega$ and a single inductor is available with reactance $100\,\Omega$ at the desired filter center frequency. In this case, a simple series or parallel RLC bandpass filter configuration will have $Q_{s,p}$ determined by the source and load resistance and the inductor value, and the bandwidth cannot be specified independently. If the inductor is used to implement a series LC filter as shown in Figure 4.19(a), the total series resistance is $100\,\Omega$ and the resulting filter has $Q_s = 1$. In

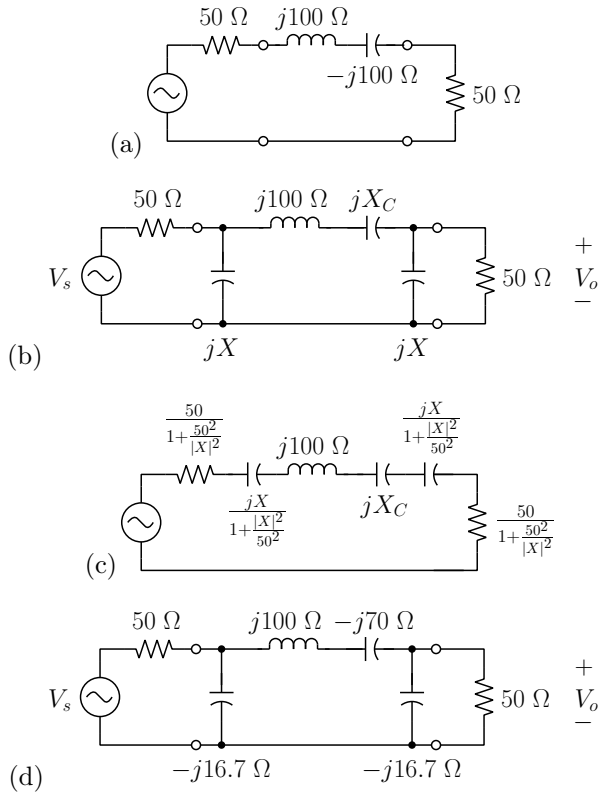


Figure 4.19: (a) With $50\,\Omega$ source and load resistances and inductor reactance $100\,\Omega$, $Q_s = 1$, and the bandwidth cannot be specified independently. (b) Shunt capacitors with reactance X ($X < 0$) have been added to transform the source and load resistances to a smaller value as shown in (c). The values given in (d) set the Q_s of the equivalent series RLC circuit (shown in (c)) to $Q_s = 10$.

Figure 4.19(b), shunt capacitors with impedance jX ($X < 0$) have been added to transform the source and load resistance to smaller values, as shown in Figure 4.19(c). The network shown in Figure 4.19(c) is similar to a series RLC filter and will have a similar frequency

response function. The resonant Q of this network can be approximated by $Q_s = 100/(2R')$, where $R' = 50/(1 + (50/|X|)^2)$. Hence, X can be chosen to set Q_s to any desired value. Suppose that the desired fractional bandwidth of the filter is $\Delta f/f_o = 0.1$ (10%). Then the required loaded $Q_s = \frac{1}{0.1} = 10$ and the total series resistance should be set to $10\,\Omega$, i.e.

$$10 = \frac{2(50)}{1 + \frac{50^2}{|X|^2}},$$

which has the solution $|X| = 50/3 = 16.7\,\Omega$. Therefore, the impedance of the shunt coupling capacitors should be set to $-j16.7\,\Omega$. The shunt capacitors are transformed to series reactances of $-j\frac{16.7}{1+(16.7/50)^2} = -j15\,\Omega$. At the resonant frequency, ω_o, the total series capacitive reactance must be equal to $100\,\Omega$ in order to resonate with the inductor, so the capacitive reactance $X_C = -70\,\Omega$. The final design is shown in Figure 4.19(d).

If the desired center frequency is $f_o = 500\,\text{MHz}$, then the values of the inductor, shunt capacitors, and series capacitor are $L = 31.8\,\text{nH}$, $C_{shunt} = 19.1\,\text{pF}$, and $C = 4.55\,\text{pF}$, respectively. The frequency response of the network with these values is shown by the solid line in Figure 4.20. For comparison, the dashed line shows the frequency response of a series RLC filter (i.e., the topology shown in 4.19(a)) with $Q_s = 10$. The series RLC filter would need an inductor and capacitor each having reactance $1000\,\Omega$ to produce a filter with $Q_s = 10$. Notice that the two curves correspond closely near the peak of the response, illustrating the validity of approximating the network shown in Figure 4.19(c) as a simple series RLC network near the resonant frequency of the network.

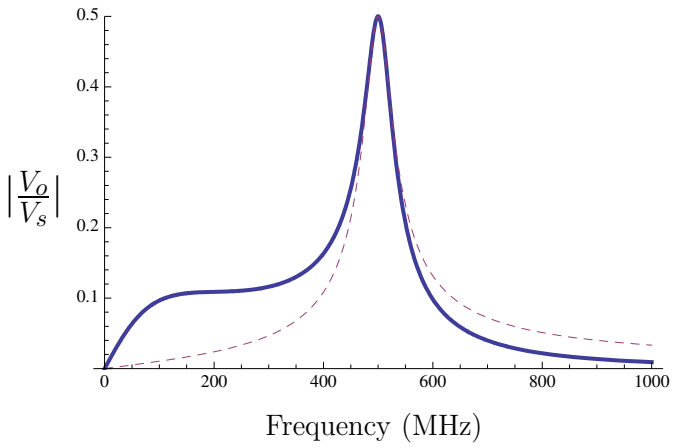


Figure 4.20: Solid line shows the frequency response of the filter shown in Figure 4.19(d), designed for a center frequency of 500 MHz. The dashed line is a plot of the function $|\frac{0.5}{1+jQ_s(\omega/\omega_o - \omega_o/\omega)}|$. The dashed line is the frequency response for a series RLC filter, as shown in Figure 4.19(a), but with the inductor and capacitor impedances equal to $\pm j1000\,\Omega$ at $f_o = 500\,\text{MHz}$. Close correspondence between the solid and dashed lines near the resonant frequency illustrates the validity of approximating the network shown in Figure 4.19(b)-(d) as a series RLC network near f_o.

4.5 Application Example - Quadrature demodulator for FM

A common application of the parallel RLC resonant circuit is found in the so-called "quadrature" demodulator for frequency modulated signals. This type of detector is implemented in a number of "receiver-on-a-chip" integrated circuits. The block diagram of a quadrature demodulator is shown in Figure 4.21. The parts within the dashed box are implemented on-chip. Typically, the multiplier is implemented with a Gilbert cell and the capacitance of C' is a relatively small value, which is easily implemented in an integrated circuit. The resonant RLC network is implemented off-chip, and is often sold under the name "quadrature coil".

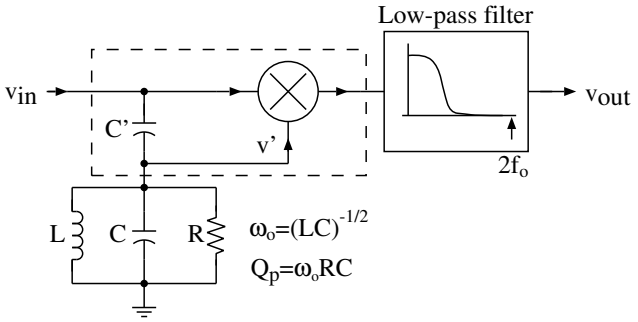


Figure 4.21: Quadrature demodulator for frequency modulated signals. The Q_p of the parallel resonant circuit determines the sensitivity and the bandwidth of the detector.

The principle of operation behind the quadrature demodulator is as follows. The frequency modulated input signal is typically provided by the output of the last IF stage of a superhet receiver. Usually, the signal will have passed through a limiter, so that the amplitude is constant. The input voltage can be written as:

$$v_{in}(t) = A\cos[(\omega_o + \Delta\omega)t] \tag{4.32}$$

where $\Delta\omega$ represents the instantaneous frequency deviation from the carrier frequency, ω_o. The deviation will be a function of time, in general, but since time variations of $\Delta\omega$ are very slow compared to the period of ω_o, we will employ a quasi-static analysis and treat $\Delta\omega$ as a (slowly varying) constant. In most applications, the carrier frequency will be equal to the last IF. The RLC circuit is tuned to resonance at ω_o. Thus, at ω_o the impedance of the RLC circuit is equal to R. If the on-chip capacitance, C', is small so that $\frac{1}{\omega_o C'} \gg R$, then the voltage v' will be shifted in phase by $\pi/2$ relative to v_{in}. Thus v' and v_{in} are in "quadrature" when the input frequency is equal to ω_o. When the frequency of the input signal deviates from ω_o, the RLC circuit looks capacitive for positive deviations, and inductive for negative deviations. This upsets the nominal quadrature relationship between v_{in} and v' and, for small frequency deviations, causes the phase difference between v_{in} and v' to vary around the nominal value of $\pi/2$ in proportion to the instantaneous frequency deviation, $\Delta\omega$. The multiplier/low-pass filter functions as a phase-detector and provides an output voltage that is proportional to the off-quadrature phase difference between v_{in} and v'. Therefore, for small frequency deviations, the output voltage is proportional to the frequency deviation.

For analytical analysis of the quadrature demodulator circuit, it is convenient to assume that the input impedance of the multiplier circuit is high, so that loading of the RLC circuit can be neglected. (Alternatively, one can assume that the finite input resistance of the multiplier has been lumped into R.) Then, using upper-case symbols to denote phasors, the voltage V' can be written in terms of the phasor input voltage, V_{in} as:

$$V' = \frac{Z_p}{Z_p + \frac{1}{j\omega C'}} V_{in} \tag{4.33}$$

where

$$Z_p = \frac{R}{1 + jQ_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})} \tag{4.34}$$

So

$$V' = \frac{R}{R + \frac{1}{j\omega C'}(1 + jQ_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}))} V_{in} \tag{4.35}$$

If C' is small enough so that $R \ll \frac{1}{\omega C'}$ then

$$V' \simeq \frac{j\omega RC'}{1 + jQ_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})} V_{in} \tag{4.36}$$

The transfer function can be re-written in terms of magnitude and phase as

$$V' = B(\omega) e^{j\theta(\omega)} V_{in}$$

where

$$B(\omega) = \frac{\omega RC'}{\sqrt{1 + Q_p^2(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})^2}} \tag{4.37}$$

and

$$\theta(\omega) = \pi/2 - \tan^{-1}[Q_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})] \tag{4.38}$$

Thus, the time-domain voltage, $v'(t)$ is given by (with $\omega = \omega_o + \Delta\omega$):

$$v'(t) = AB(\omega)\cos[\omega t + \theta(\omega)] \tag{4.39}$$

The multiplier forms the product $v_{in}v'$, so the signal at the output of the multiplier is:

$$v_{in}(t)v'(t) = A^2 B(\omega)\cos[\omega t]\cos[\omega t + \theta(\omega)] \tag{4.40}$$

Taking the nominal phase-shift, $\pi/2$, in the expression for $\theta(\omega)$ into account, we can write

$$v_{in}(t)v'(t) = -A^2 B(\omega)\cos[\omega t]\sin[\omega t - \tan^{-1}[Q_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})] \tag{4.41}$$

The sine-cosine product yields a "double frequency" term, and a "difference frequency" term. The lowpass filter removes the double frequency term, so the output of the system can be written:

$$v_o = \mathrm{LPF}[v_{in}(t)v'(t)] = \frac{1}{2}A^2 B(\omega)\sin[\tan^{-1}(Q_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}))] \tag{4.42}$$

Employing the trigonometric identity $\sin[\tan^{-1} x] = \frac{x}{\sqrt{1+x^2}}$ the output voltage can be written as

$$v_o = \frac{1}{2}A^2 B(\omega)\frac{Q_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})}{\sqrt{1 + Q_p^2(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})^2}}. \tag{4.43}$$

Inserting the expression for $B(\omega)$, we find:

$$v_o = \frac{1}{2}A^2 \omega RC' \frac{Q_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})}{1 + Q_p^2(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})^2}. \tag{4.44}$$

Now, with $\omega = \omega_o + \Delta\omega$

$$\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega} = \frac{\omega_o + \Delta\omega}{\omega_o} - \frac{\omega_o}{\omega_o + \Delta\omega} = \frac{2\Delta\omega}{\omega_o + \Delta\omega} + \frac{\Delta\omega^2}{\omega_o(\omega_o + \Delta\omega)}. \tag{4.45}$$

For small frequency deviations such that $\Delta\omega \ll \omega_o$:

$$\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega} \simeq \frac{2\Delta\omega}{\omega_o}. \tag{4.46}$$

So, for small frequency deviations we have

$$v_o \simeq \frac{1}{2}A^2 \omega_o RC' \frac{Q_p \frac{2\Delta\omega}{\omega_o}}{1 + Q_p^2 \frac{4\Delta\omega^2}{\omega_o^2}} = \frac{1}{2}A^2 Q_p \frac{C'}{C} \frac{Q_p \frac{2\Delta\omega}{\omega_o}}{1 + Q_p^2 \frac{4\Delta\omega^2}{\omega_o^2}}. \tag{4.47}$$

This expression has the form:

$$v_o \sim \frac{x}{1 + x^2}$$

where $x = 2Q_p\Delta f/f_o$. This function is plotted in Figure 4.22.

Some quadrature detector implementations use a multiplier with "limiting" inputs, i.e. one or both of the inputs is driven into saturation so that the multiplier output is insensitive to changes in the amplitude of the input signals. If the v' input is saturated, then the output becomes independent of the amplitude of v'. In this case, equation 4.40 must be modified by replacing the term $A^2B(\omega)$ with a constant. In this case, the output can be written:

$$v_o = K\frac{Q_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})}{\sqrt{1 + Q_p^2(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})^2}}. \tag{4.48}$$

Using the same approximations that were used before, we obtain:

$$v_o \simeq K\frac{Q_p \frac{2\Delta\omega}{\omega_o}}{\sqrt{1 + Q_p^2 \frac{4\Delta\omega^2}{\omega_o^2}}} \tag{4.49}$$

which has the form:

$$v_o \sim \frac{x}{\sqrt{1 + x^2}}$$

with $x = 2Q_p\Delta f/f_o$. This function is plotted in Figure 4.23.

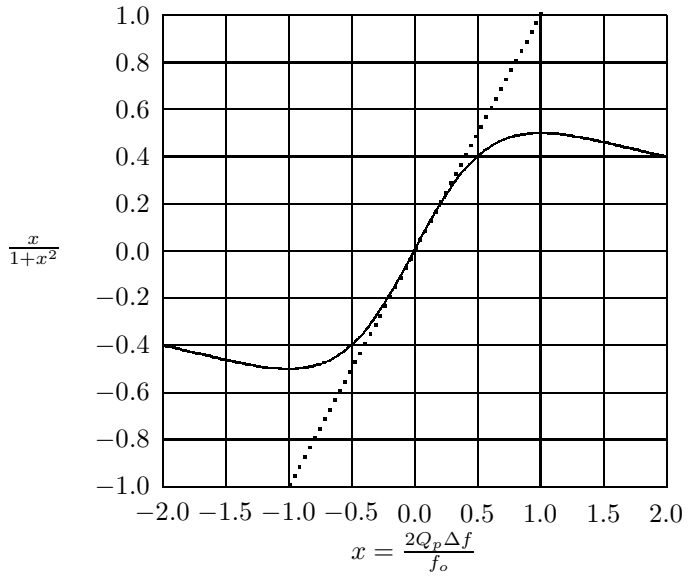


Figure 4.22: Detector output as a function of frequency deviation. This curve is referred to as the demodulator's "S-curve". Notice that the region where the output is approximately proportional to the frequency deviation is limited $|2Q_p\Delta f/f_o| \ll 1$. The dotted line shows an ideal linear response with the same slope as the actual response at $x = 0$.

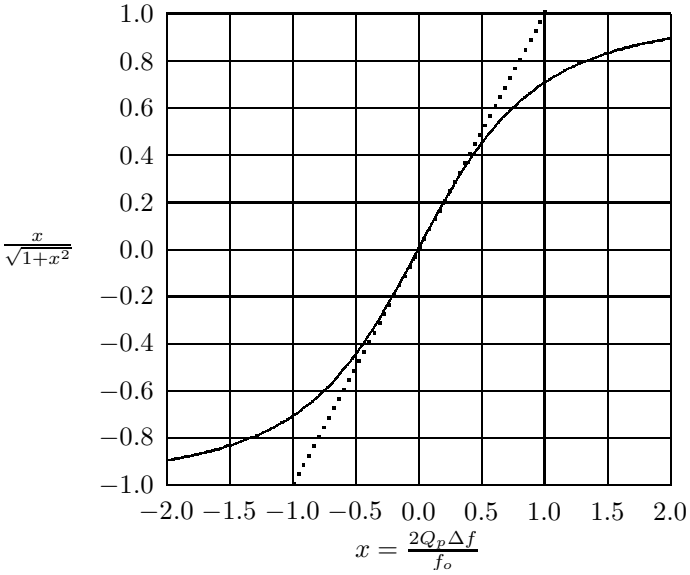


Figure 4.23: Output voltage verses normalized frequency deviation for a quadrature demodulator employing a saturated input for v'. The dotted line shows an ideal linear response with the same slope as the actual response at $x = 0$.

In either case (ideal multiplier, or multiplier with saturated inputs), if the frequency deviation is small enough, i.e. if $2Q_p\Delta\omega/\omega_o \ll 1$, then the output voltage is proportional to the instantaneous frequency deviation:

$$v_o \sim \frac{Q_p}{f_o}\Delta f, \quad \Delta f \ll f_o \text{ and } 2Q_p\Delta f \ll f_o \tag{4.50}$$

The coefficient in front of Δf determines the sensitivity of the demodulator. It is apparent that to obtain larger output voltage for a given deviation one would like to choose a higher Q_p for the resonant RLC circuit. On the other hand, the maximum Q_p that can be employed is limited by the requirement that $2Q_p\Delta f_{max}/f_o \ll 1$ (where Δf_{max} is the peak frequency deviation of the modulated input signal). Recall that the -3 dB bandwidth of an RLC filter can be written in terms of the Q, i.e. $BW = f_o/Q_p$, so the constraint on Q_p can be rewritten in terms of a constraint on the bandwidth of the RLC circuit: $BW \gg 2\Delta f_{max}$. In other words, the -3 dB bandwidth of the RLC circuit should be large compared to the deviation of the input signal in order for the demodulator to have a linear response.

A useful practical constraint for design purposes is to require that the demodulator's response characteristic should not deviate from perfect linearity by more than 1% over the whole range of deviation, i.e. for $-\Delta f_{max} < \Delta f < \Delta f_{max}$. If the demodulator is implemented with an ideal multiplier, this requirement leads to $x < 0.100$. The largest Q_p that will satisfy this constraint is:

$$Q_p = 0.1\frac{f_o}{2\Delta f_{max}} \tag{4.51}$$

If a multiplier with saturated inputs is used to implement the detector, then the departure from linearity will be held to the 1% limit if $x < 0.142$. The largest Q_p that satisfies this constraint is:

$$Q_p = 0.142\frac{f_o}{2\Delta f_{max}} \tag{4.52}$$

The extent to which the demodulator's characteristic departs from linearity will determine the amount of distortion in the demodulated signal's waveform. For some applications, such as demodulation of frequency-shift-keyed (FSK) data signals, low distortion is not particularly important. In such cases a larger deviation from linearity can be tolerated in return for larger output voltage from the detector.

4.6 References

1. Terman, Frederick Emmons, *Radio Engineers Handbook*, McGraw Hill, 1943.

4.7 Homework Problems

1. The circuit shown in Figure 4.24 is usually a good model for a parallel LC circuit implemented with real, lossy components. The element values are $C = 800$ pF, $L = 15\,\mu$H, $r = 1\,\Omega$, $R = 10\,\text{k}\Omega$. Use series to parallel transformations to transform this circuit into an equivalent parallel RLC circuit and find the approximate resonant frequency and Q_p of the equivalent circuit.

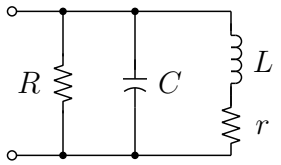


Figure 4.24: Resonant circuit formed from lossy capacitor (modeled by R and C) and a lossy inductor (modeled by r and L).

2. Two series RLC circuits are connected in series to form a new resonant circuit. Denote the elements of the individual resonant circuits by (R_1, L_1, C_1) and (R_2, L_2, C_2), respectively. The series resonant frequency of each of the two circuits is the same and is denoted by ω_o. The Q's of the two circuits are Q_1 and Q_2.

   (a) What is the resonant frequency of the series combination of the two circuits?

   (b) Find an expression for the Q of the overall circuit. Express your result in terms of Q_1, Q_2, R_1 and R_2 only.

   (c) Is the statement below TRUE or FALSE? You must justify your answer to receive full credit.

$$\min(Q_1, Q_2) \leq Q \leq \max(Q_1, Q_2) \tag{4.53}$$

3. One property of a parallel resonant circuit is that the current in the reactive components can be much larger than the applied current. This must be considered when choosing the current ratings of components such as capacitors and inductors (especially in high power devices such as transmitters). Show that the peak current through the inductor or capacitor in a parallel resonant circuit is $Q_p I_i$ at resonance, where I_i is the current supplied to the entire resonant circuit

4. One property of a series resonant circuit is that the voltage across the reactive components can be much larger than the applied voltage. This must be considered when choosing the voltage ratings of components such as capacitors and inductors (especially in high power devices such as transmitters). Show that the peak voltage across the capacitor in a series resonant circuit is $Q_s V_i$ at resonance where V_i is the voltage across the entire resonant circuit. Note that the peak voltage across the inductor will be the same.

5. Consider the circuit shown in Figure 4.25. The current source has constant amplitude and frequency f_c, and it drives a bandpass filter consisting of a lossy inductor in parallel with a variable capacitor and a resistor. You may assume that any capacitance associated with the inductor has been incorporated into the variable capacitor indicated in the schematic. The variable capacitor C can be set to any value in the range $36 - 365\,\text{pF}$, and $r = 10\,\Omega$, $R = 100\,\text{k}\Omega$. Suppose that the frequency of the current

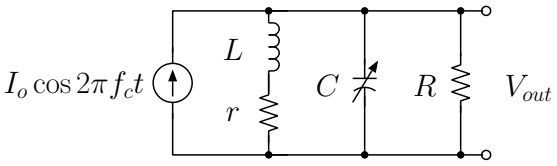


Figure 4.25: LC filter with variable center frequency.

source can be adjusted to any frequency in the range 540-1700 kHz. For a given value of the source frequency, f_c, the variable capacitor will be tuned to maximize the output voltage. Approximate the filter as a parallel RLC circuit to answer the following questions:

   (a) Specify a single value of L that would allow the variable capacitor to tune the filter (i.e. maximize the output voltage) to any frequency in the range 540-1700 kHz. Give your result in μH.

   (b) With the value of L that was determined in part (a), determine the approximate 3dB bandwidth of the filter when $f_c = 540$ kHz and C is adjusted to maximize the output voltage.

   (c) Same as part (b) but for $f_c = 1700$ kHz.

6. The circuit in Figure 4.26 is a model for the operation of a "ferrite loopstick" antenna that is commonly employed in AM broadcast band radios. The antenna is tuned to resonance by a capacitor C which is usually adjustable to allow the circuit to cover the entire broadcast band. This circuit also performs the function of the preselector. The voltage source V_s represents the emf induced in the coil as a result of an incident electromagnetic wave with frequency ω. The resistance r represents the losses in the coil and R represents the input impedance of the following stage.

   (a) Find an exact expression in terms of r, R, L, and C for the frequency where the output voltage is a maximum (the resonance frequency).

   (b) Suppose that $V_s = (1\,mV)\,\cos[\,2\pi\,f_c\,t\,]$, $L = 100\mu H$, $R = 100k\,\Omega$, and $r = 6\,\Omega$. Find the range of values that C must cover in order for the circuit to tune the AM broadcast band (540-1700 kHz). Note: For this purpose you need an expression for the value of C that maximizes the voltage response at a given frequency. An approximate analysis is acceptable, but be sure to carefully state and justify your assumptions.

   (c) Now suppose that the source frequency, f_s, is swept from 540 to 1700 kHz. The zero-to-peak value of the source voltage is held constant at 1 mV as the frequency

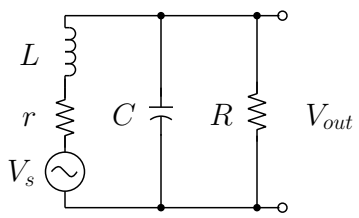


Figure 4.26: Ferrite loopstick antenna model.

is swept. Also suppose that the circuit is tuned to follow the frequency of the source so that the output voltage is always maximized. Plot the output voltage as a function of frequency.

7. Consider a parallel resonant circuit that is used as the preselector of an AM broadcast band receiver. The inductor used in the resonant circuit is near the peak of its "Q-versus-frequency" curve in the band of frequencies from 550-1600 kHz, and hence the Q_L of the inductor can be assumed to be constant over the band. Assume that $Q_L = 100$ and that the capacitor is lossless.

   (a) What will the 3 dB bandwidth of the preselector be when the preselector is tuned to 550 kHz? How about 1600 kHz?

   (b) Now consider a different situation where the inductor is chosen such that the skin effect and distributed capacitance can be neglected for the frequencies of interest, i.e., the inductor is operated at frequencies well below the peak of its "Q-versus-frequency" curve. Suppose the inductor has $Q_L = 50$ at 550 kHz. What will the 3 dB bandwidth of the preselector be at 550 and 1600 kHz?

8. In a series RLC circuit that is resonant at 1150 kHz it is found that when the frequency differs from resonance by 15 kHz, the current drops to 0.53 of the current at resonance, for the same applied voltage. Determine the Q_S of the circuit.

9. Consider the circuit shown in Figure 4.27 where $R_S = R_L = 50\,\Omega$, $L = 48\,\mu$H, $C = 10$ pF:

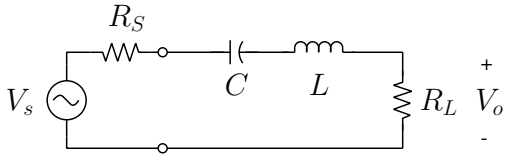


Figure 4.27: Series resonant circuit

   (a) Find the resonant frequency where maximum power would be delivered to the 50 Ω load resistor. Express your result in MHz.

(b) Suppose that the time-averaged power delivered to the 50 Ω load resistor at resonance is 1000 W. Find the peak value of the voltage across the capacitor.

(c) Approximately how far off of resonance would the source frequency need to be moved in order to cause the average power delivered to the load resistor to drop by 6 dB? Express your result in kHz.

Note: For small frequency shifts, $\Delta\omega$, around the resonant frequency, ω_o, the following approximation is useful:

$$\frac{\omega_o + \Delta\omega}{\omega_o} - \frac{\omega_o}{\omega_o + \Delta\omega} \approx 2\frac{\Delta\omega}{\omega_o} \tag{4.54}$$

10. Consider the circuit shown in Figure 4.28 and its Thevenin equivalent circuit.

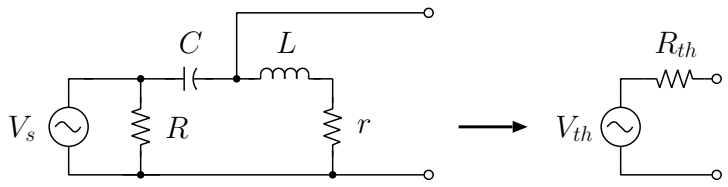


Figure 4.28: Series resonant circuit with Thevenin equivalent circuit.

(a) If $V_s = 2\,\text{V}$, $R = 1000\,\Omega$, $C = 1000\,\text{pF}$, $L = 10\,\mu\text{H}$, $r = 1\,\Omega$, find the (non-zero) frequency, ω_o, where the indicated Thevenin equivalent circuit would be valid. In other words, find the frequency where the Thevenin impedance is purely resistive. Clearly state any assumptions or approximations that you make.

(b) Find R_{th} and V_{th} at the frequency found in part (a). For V_{th}, specify magnitude and phase (in degrees).

11. Shown in Figure 4.29 are two different ways of coupling a voltage source to a load resistor using a resonant circuit for the purpose of providing a bandpass output voltage response. The values are: $V_s = 10V$ (peak), $R_s = 500\,\Omega$, $R_L = 500\,\Omega$, $L = 1\,\mu H$, C = 500 pF.

Compute the following for each circuit:

(a) The output voltage at resonance.

(b) The 3dB bandwidth of the transfer function V_o/V_s. Express your result in MHz.

(c) The upper and lower 3 dB frequencies, i.e., give the actual frequencies (in MHz). The difference between these frequencies will be the 3dB bandwidth that you found in part 11b.

12. Consider the capacitive transformer drawn in Figure 4.30.

Under certain conditions this circuit will approximately transform the resistance R to a new resistance n^2R in parallel with an effective capacitance, C_p, where $n = 1 + \frac{C_1}{C_2}$.

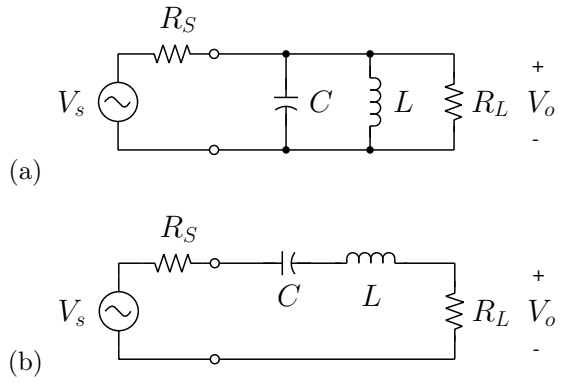


Figure 4.29: Voltage source coupled to load resistor with (a) parallel LC circuit and (b) series resonant circuit to provide a bandpass voltage transfer function V_o/V_s.

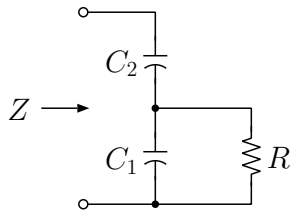


Figure 4.30: Capacitive transformer

(a) Under what condition(s) will this occur?

(b) Under this condition, give an expression for C_p.

13. The circuit shown in Figure 4.31 is a capacitive transformer with resonating inductance L. Suppose that

$$\begin{aligned} R &= 50\,\Omega \\ C_1 &= 3183\,\text{pF} \\ C_2 &= 3183\,\text{pF} \end{aligned}$$

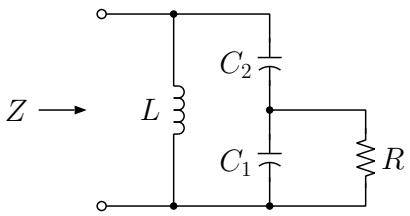


Figure 4.31: Resonant capacitive transformer

Use parallel-series and series-parallel transformations, with appropriate approximations, and find:

(a) The inductance, L, required to resonate the circuit at 10 MHz. At resonance, the impedance Z will be purely real.

(b) The input impedance, Z at 10 MHz.

(c) The Q of the circuit can be approximated by the Q_p of the equivalent parallel RLC circuit (at 10 MHz). Find the approximate Q.

To save time, note that the reactances of C_1 and C_2 have magnitude 5 Ω at 10 MHz.

14. Consider in Figure 4.32 the small signal model for a tuned-output amplifier:

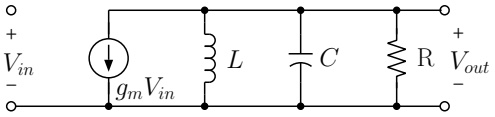


Figure 4.32: Small signal model for tuned-output amplifier.

Suppose that N of these stages are cascaded. Show that the 3dB bandwidth of the cascade will be given by

$$BW = BW^1\sqrt{2^{1/N} - 1}\ (Hz) \tag{4.55}$$

where BW^1 is the 3 dB bandwidth of a single stage

$$\begin{aligned} BW^1 &= \frac{f_0}{Q} \\ &= \frac{1}{2\pi RC} \end{aligned} \tag{4.56}$$

15. Consider the circuit shown in Figure 4.33. Define the following quantities:

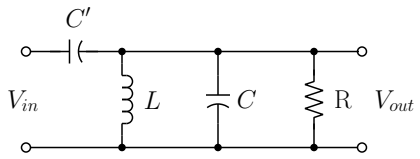


Figure 4.33: Phase shifter based on parallel RLC circuit.

$$\omega_o = \frac{1}{\sqrt{LC}} \quad \text{and} \quad Q_p = \omega_o RC \tag{4.57}$$

(a) Find an expression for the phase shift between V_{out} and V_{in}, i.e., find an expression for the phase angle of the transfer function $T(\omega) = V_{out}(\omega)/V_{in}(\omega)$.

(b) Show that if the reactance of the capacitor C' is much larger than R, then the expression for the phase shift reduces to

$$\theta = \pi/2 - \tan^{-1}[Q_p(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})] \tag{4.58}$$

(c) Show that the phase shift is approximately a linear function of frequency for frequencies near the resonant frequency of the tuned circuit.

16. A 100 pF capacitor and 2.53 μH inductor are used to form a series resonant circuit at 10 MHz. The capacitor is lossy and has a component Q of 100 at 10 MHz. The inductor is also lossy and has a component Q of 80 at 10 MHz.

(a) Find the Q_s of the resonant circuit formed by the series combination of the lossy capacitor and inductor.

(b) Find the impedance of the series resonant circuit at 10 MHz.

(c) Suppose that it is desired to couple a 10 Ω source to a 10 Ω load through a bandpass filter formed either by:

i. a series resonant circuit or

ii. a parallel resonant circuit.

In either case, the resonant circuit would be formed by the lossy L and C considered in parts 16a and 16b. Sketch a schematic showing how the source should be coupled to the load if the smallest possible bandwidth is desired.

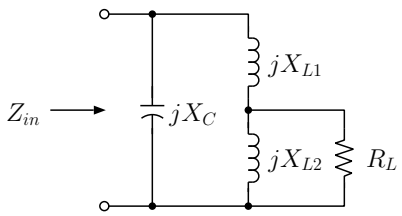
(d) Estimate the bandwidth (in MHz) of the filter that you sketched in part 16c.

17. A series LC circuit consisting of an ideal lossless inductor L and ideal lossless capacitor C is used as a bandpass filter to couple a source with source impedance $R_S = 50\,\Omega$ to a load $R_L = 50\,\Omega$. The resonant frequency of the filter is $f_o = 100$ MHz, and the values are selected such that the -3 dB bandwidth of the filter is 10 MHz.

(a) Sketch the system, including the source, the LC filter, and the load and indicate the values for L and C in nH and pF, respectively.

(b) Now, replace the inductor and capacitor with lossy components that have the same inductance and capacitance as the original components but with component Q's that are equal to 40 at the resonant frequency, i.e. $Q_C = Q_L = 40$ at 100 MHz. Calculate the -3 dB bandwidth of the filter implemented using the lossy components.

18. The circuit shown below is used to transform a load resistance, R_L, to a new value, R_{in}, at the resonant frequency of the network. Away from resonance, the circuit provides a bandpass filter response.



(a) The load resistance $R_L = 50\,\Omega$. At the resonant frequency assume that $X_{L1} = X_{L2} = 5\,\Omega$. Determine the value of X_C required to resonate the circuit (i.e. to make the input impedance Z_{in} purely resistive).

(b) For the value of X_C found in part a. determine Z_{in}.

(c) The Q of the resonant circuit can be approximated by Q_p of the equivalent parallel RLC circuit obtained by applying parallel<=>series transformations. Find the approximate Q of the network.

19. A 50 Ω source is coupled to a 50 Ω load through a bandpass filter consisting of a series LC circuit which is resonant at 20 MHz.

(a) Sketch the system, including the source, the filter, and the load.

(b) Design the filter (i.e., find L and C) so that the -3 dB bandwidth of the filter is 5 MHz.

(c) Now suppose the inductor is lossy and has finite component Q, denoted by Q_L. (You may assume that the inductance is the same as determined in part b.) What is the minimum acceptable Q_L if the attenuation caused by the lossy L at 20 MHz is to be smaller than 2 dB?

20. Design an LC circuit consisting of an ideal lossless inductor L and ideal lossless capacitor C to be used as a bandpass filter to couple a source with source impedance $R_S = 100\,\Omega$ to a load $R_L = 100\,\Omega$. Assume that a limited range of inductance values is available such that $10\,\text{nH} \leq L \leq 100\,\text{nH}$.

    (a) Determine the topology of the LC filter and the values for L and C if the filter is to have center frequency $f_0 = 100\,\text{MHz}$ and *the smallest possible bandwidth*. Sketch the system, including the source, the LC filter, and the load. Specify the values for L and C in nH and pF, respectively.

    (b) Repeat part a for the center frequency $f_0 = 1\,\text{GHz}$.

21. An LC circuit consisting of an ideal lossless inductor L and ideal lossless capacitor C is to be used as a bandpass filter to couple a source with source impedance R_S to a load R_L. Assume that $R_S = R_L = R$. Denote the resonant frequency of the filter by ω_o, and denote the reactance of the inductor and capacitor at ω_o by $\pm X_o$.

    (a) If $X_o > R$, sketch the system (source, load, and LC filter) that will provide the smallest possible bandwidth.

    (b) Consider the system that you sketched in part a., and denote the loaded Q of the system by Q_o. The -3 dB bandwidth of the system is then $BW_o = \frac{\omega_o}{Q_o}$. Now, suppose that the system is modified by replacing the lossless inductor with a lossy inductor. Denote the component Q of the lossy inductor at ω_o by Q_L. The inductance of the lossy inductor is the same as the inductance of the original lossless component. Denote the bandwidth of the modified circuit by BW'. Find an expression for BW' that involves only BW_o, Q_o, and Q_L.

    (c) If $Q_o = 10$ and $Q_L = 40$, by what percentage is the bandwidth of the filter increased because of the loss in the inductor?

    (d) Let P_L denote the power delivered to the load at ω_o with the lossless inductor in the system, and let $P_L^{'}$ denote the power delivered to the load at ω_o with the lossy inductor in the system. Find an expression for the power ratio, $P_L^{'}/P_L$. Express your result only in terms of Q_o and Q_L.

    (e) If $Q_o = 10$ and $Q_L = 40$, determine the power loss, in dB, of the filter at the resonant frequency.

22. An LC circuit consisting of an ideal lossless inductor L and ideal lossless capacitor C is to be used as a bandpass filter to couple a source with source impedance R_S to a load R_L. Assume that $R_S = R_L = R$. Denote the resonant frequency of the filter by ω_o, and denote the reactance of the inductor and capacitor at ω_o by $\pm X_o$.

    (a) If $X_o < R$, sketch the system (source, load, and LC filter) that will provide the smallest possible bandwidth.

(b) Consider the system sketched in part a., and denote the loaded Q of the system by Q_o. Now, suppose that the system is modified by replacing the lossless inductor with a lossy inductor and the lossless capacitor with a lossy capacitor. Denote the component Q's of the lossy inductor and capacitor at ω_o by Q_L and Q_C, respectively. The inductance (capacitance) of the lossy inductor (capacitor) are the same as the inductance and capacitance of the original lossless components. Denote the new loaded Q of the system by Q'_o. Find an expression for Q'_o that involves only Q_o, Q_L, and Q_C. You may assume that $Q_L \gg 1$ and $Q_C \gg 1$.

23. Consider the single-resonator bandpass filter shown in Figure 4.34.

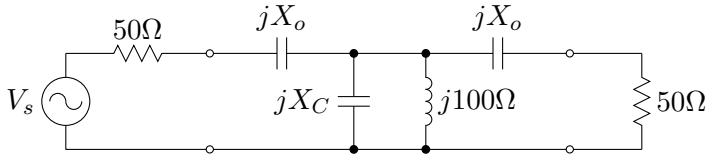


Figure 4.34: Single-resonator bandpass filter.

(a) Suppose that the inductor has reactance $X_L = 100\ \Omega$ at the center frequency of the filter. Find the reactances of the coupling capacitors, $X_o = -1/(\omega_o C_o)$, and the reactance of the resonator capacitor, $X_C = -1/(\omega_o C)$, such that the filter has a fractional -3 dB bandwidth of 5%. Hint: use series to parallel transformations to convert the circuit to a parallel RLC circuit.

(b) Find L, C, C_o that will place the center frequency of the filter at $f_o = 500$ MHz.

(c) Using the values that you found in part b, plot the voltage frequency response of the filter. Feel free to use your favorite circuit simulation software.

(d) Suppose that you build the filter using an inductor that has $Q_L = 50$ at f_o. Assume that the capacitors are lossless. Estimate the filter loss at f_o. Express your result in dB.

24. The system shown in Figure 4.35 consists of a resistive source and load coupled with a bandpass filter. Design the filter to have fractional bandwidth $\frac{\Delta f_{3dB}}{f_o} = 0.04$. The filter

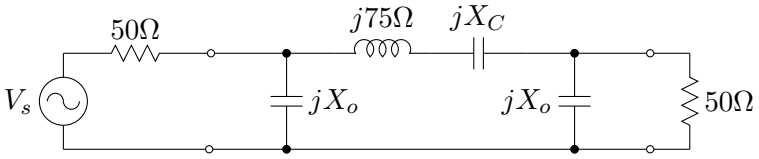


Figure 4.35: Series-resonator bandpass filter.

uses a single inductor with reactance $j50\,\Omega$ at f_o. The reactances of the capacitors at f_o are denoted by X_o and X_C.

(a) Find numerical values for X_o and X_C. Note that $X_C < 0$ and $X_o < 0$.

(b) Find the component values for the inductor and capacitors (L, C_C, C_o) that will place the center frequency of the filter at $f_o = 50$ MHz.

(c) Using the values that you found in part b, plot the voltage frequency response of the filter. Feel free to use your favorite circuit simulation software.

(d) Suppose that you build the filter using an inductor that has $Q_L = 50$ at f_o. Assume that the capacitors are lossless. Estimate the filter loss at f_o. Express your result in dB.

25. Consider a simple single-resonator bandpass filter consisting of a series LC branch in series with the source and load, or a parallel LC branch in parallel with the source and load. Recall that the frequency response function of such a filter can be written in terms of the loaded Q and the center frequency, ω_o, as follows:

$$F(\omega) = \frac{1}{1 + jQ(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega})}.$$

(a) Find an expression for the bandwidth, Δf_{-A}, at attenuation level A, where A is the attenuation expressed in dB. Express your result in terms of the power ratio $\alpha = 10^{A/10}$, the loaded Q of the filter, and the center frequency f_o.

(b) Use your result from part (a) to anwer this question: A single-resonator LC bandpass filter has center frequency $f_o = 10\,\text{MHz}$ and loaded $Q_{s,p} = 10$. Find the -20 dB bandwidth, $\Delta f_{-20\,\text{dB}}$, of the filter.

(c) For the filter described in part b, find the lower -20 dB frequency of the filter frequency response. Express your result in MHz.

Chapter 5

Oscillators

5.1 Introduction

This chapter will describe the methods used to analyze and design oscillators for applicatons in transmitters and receivers. We will employ small-signal linear analysis techniques to determine whether or not a particular circuit will oscillate, and a simplified nonlinear model will be used to estimate the amplitude of the oscillations in a circuit based on a BJT. Numerical simulations will be used to illustrate the characteristics of realistic oscillator circuits, and the results will be compared to the analytical predictions.

For small signal analysis it is convenient to think of oscillators as unstable feedback systems. An unstable system is one in which an initially small excitation or disturbance produces an output that grows in time due to constructive, or positive, feedback. A block diagram of a simple feedback system is shown in Figure 5.1.

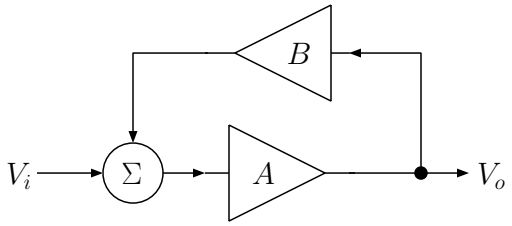


Figure 5.1: A simple feedback system

The voltage transfer function for this system is

$$\frac{V_o}{V_i} = \frac{A(j\omega)}{1 - A(j\omega)B(j\omega)} \tag{5.1}$$

where the quantity $A_{lo}(\omega) = A(j\omega)B(j\omega)$ is called the *open loop gain* of the system — sometimes shortened to *loop gain*. The subscript "o" is used to indicate that the loop gain is computed assuming *small-signal* operation of the active devices. Note carefully that the loop gain is the gain obtained by opening the feedback loop and taking the output at the point

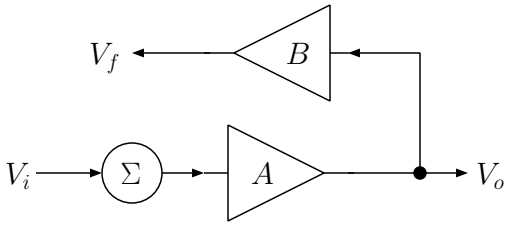


Figure 5.2: Loop gain at summing junction

where the loop was opened. For example, if the loop is opened at the summing junction, then the loop gain is V_f/V_i as shown in Figure 5.2.

Suppose that the loop gain is equal to 1 at some frequency ω_o, i.e., $A_{lo}(\omega_o) = 1$. Then the voltage transfer function in Figure 5.1 is singular (infinite), which can be interpreted as finite output for zero input. In other words, the circuit is a potential source of radio frequency energy at the frequency where the loop gain is 1, even in the absence of any input excitation V_i.

The condition for steady-state oscillation ($A_{lo}(\omega_o) = 1$) is intuitively satisfying - when the loop gain is 1, a sinusoidal excitation presented to the input of the circuit traverses the feedback loop and appears back at the input with the same amplitude and phase that it started with. This re-circulation of the disturbance proceeds indefinitely, with the circuit "oscillating" in a steady state. In practice the *small-signal* loop gain is set to a value somewhat larger than 1. This means that the disturbance is amplified after each pass through the loop, and the output grows as the disturbance passes repeatedly through the loop. In most radio frequency oscillators the loop gain is equal to 1 (or some real number larger than 1) at one particular frequency. At other frequencies the amplitude may be less than or greater than 1, but the phase angle is non-zero. This means that only one frequency component can travel around the loop with no phase shift. Only this frequency component will be amplified and grow to become a steady-state oscillation.

We shall see later in this chapter that even when the condition for oscillation is satisfied at only one frequency, it is possible to have a non-sinusoidal output. However, this is due to nonlinear effects in the amplifying devices. As noted above, practical oscillator circuits are designed so that the *small-signal* loop gain is larger than one at the desired frequency of oscillation, ω_o. This ensures that even thermal-noise-level signals will provide enough excitation to cause oscillation to start and grow at ω_o. Some mechanism must be built into the circuit to limit the amplitude of the oscillation to a finite value. The loop gain must be decreased as the oscillation amplitude grows so that the loop gain can eventually settle down to the steady-state value of 1.0. In so-called self-limiting oscillators, the amplitude of the oscillation eventually becomes large enough to begin to saturate the active device, effectively reducing the gain of the device, and also reducing the loop gain. Oscillation amplitude stabilizes at the amplitude where the gain of the active device is reduced just enough to set the loop gain to 1.0. Since self-limiting oscillators rely on driving the active device to levels where nonlinearity is important, they often produce non-sinusoidal outputs with significant harmonic content.

In some cases, it is desired to limit the amplitude of the oscillation at a level smaller than that required to saturate the active device. One mechanism for doing this was implemented

by Hewlett and Packard in their first commercial product, an audio oscillator. They used a small light bulb as a resistor to set the voltage gain of the amplifier in the oscillator. As oscillation amplitude grows, the light bulb filament heats up, and the resistance of the filament increases, reducing the gain of the amplifier. The thermal time constant of the filament was much longer than the period of the oscillation, so that the filament acts as a linear resistor, with resistance proportional to the oscillation amplitude. This mechanism allowed Hewlett and Packard to create an audio oscillator with a nearly perfect sinewave output. The oscillator exhibited an extremely low level of harmonic distortion in the output because the amplitude of oscillation was limited at a level well within the linear range of the amplifier.

In any case, some initial disturbance is necessary to start the oscillations. As long as the small-signal loop gain is set equal to a value greater than 1 at the potential frequency of oscillation, the thermal noise voltage that is always present in electrical circuits and/or the turn-on transient which results when power is applied to the circuit will provide the initial disturbance that starts the oscillations.

5.2 Oscillator Analysis using Loop Gain

We will now proceed to analyze one type of oscillator circuit using the loop gain approach. The analysis proceeds as follows. First the feedback loop must be identified and the loop gain computed. Then the condition for oscillation is applied to the loop gain. Oscillation occurs if there is some frequency where the loop gain has magnitude 1 and phase angle equal to zero. The combination of these two constraints is called the *Barkhausen Criterion* for oscillation. The two conditions can be written as

$$arg[A_{lo}(\omega_o)] = 0 \tag{5.2}$$

$$|A_{lo}(\omega)|_{\omega=\omega_o} = 1 \tag{5.3}$$

In practice we are usually able to apply condition (5.2) to solve for the potential frequency of oscillation, ω_o. Then applying condition (5.3) will determine how much gain the amplifier must have in order to make the loop gain equal to 1 at ω_o. The gain value determined by equation (5.3) is the value required to support steady-state oscillation (oscillation amplitude neither growing nor decaying). For practical designs, the amplifier gain is set to a value somewhat higher than the value determined by equation (5.3) to ensure that oscillations will reliably start and grow when the circuit is powered up.

Circuits with the topology shown in Figure 5.3a are commonly employed as oscillators. The active device could be a BJT or an FET. This circuit can be analyzed as a feedback loop. The circuit is redrawn in Figure 5.3b to explicitly show that the feedback from output to input is through the element Z_3. The loop gain is easily computed for circuits of the type represented by Figure 5.3. For small-signal analysis we can model the transistor as shown in Figure 5.4. Note that this is a simplified version of the hybrid-pi model for the transistor (see Appendix A). The passive elements of the model (e.g., r_π, C_π, C_μ, etc.) can be lumped into the external impedances Z_1, Z_2, Z_3. To find the open loop gain of this circuit, we break the loop shown in Figure 5.3b at a convenient point and terminate that point in the impedance that it sees when the loop is closed. The termination is required so that the open-loop circuit operates under the same conditions as when the loop is closed. This procedure will give the value of the loop gain that is appropriate for closed loop conditions.

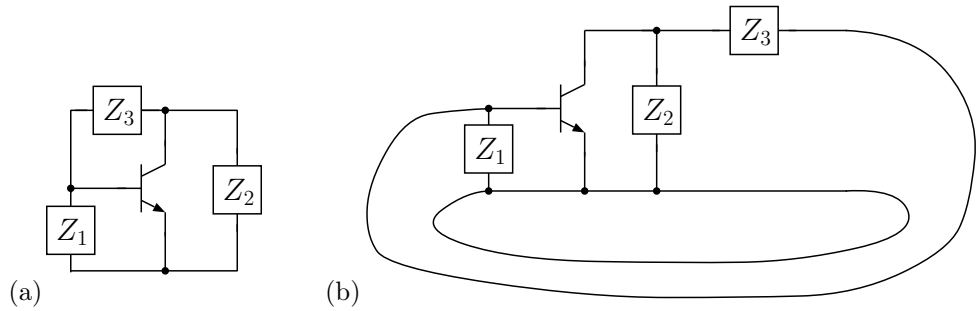


Figure 5.3: (a) Topology of one class of oscillator circuits. (b) Same as (a), redrawn to show the feedback path from output to input through Z_3.

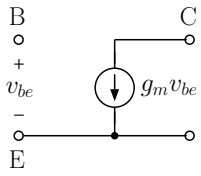


Figure 5.4: Simplified hybrid-pi model.

For example, suppose that the loop is broken to the right of Z_3 in Figure 5.3b. Since the output of Z_3 normally looks into Z_1 when the loop is closed, we terminate the loop with Z_1 as shown in Figure 5.5. Then the loop gain is computed by exciting the circuit at the input

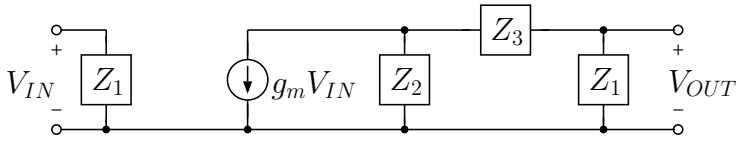


Figure 5.5: Feedback loop terminated with Z_1

to the opened loop and computing the output across Z_1:

$$A_{lo} = \frac{V_{OUT}}{V_{IN}} = \frac{-g_m Z_1 Z_2}{Z_1 + Z_2 + Z_3} \tag{5.4}$$

Some useful insights can be gained if we assume for the moment that Z_1 and Z_2 are purely reactive, i.e., $Z_1 = jX_1$, $Z_2 = jX_2$. We allow Z_3 to have a non-zero (positive) real part. Then

$$A_{lo} = \frac{g_m\, X_1\, X_2}{Z_3 + j(X_1 + X_2)} \tag{5.5}$$

For oscillation to occur, the phase angle of A_{lo} must be zero at some frequency. Now there are four possibilities for the signs of X_1 and X_2. Define $Z_s = Z_3 + j(X_1 + X_2) = |Z_s|e^{j\theta_{Z_s}}$. Then for each possible choice of sign on X_1 and X_2, there is a requirement on the phase angle of Z_s (denoted by θ_{Z_s}) in order for the overall phase angle of A_{lo} to be zero. These constraints are summarized in Table 5.1.

Table 5.1: Constraints for Overall Phase Angle of A_{lo} to be Zero

(1)	$X_1 > 0$ (inductor),	$X_2 > 0$ (inductor)	$\Rightarrow$	$\theta_{Z_s} = 0$
(2)	$X_1 < 0$ (capacitor),	$X_2 < 0$ (capacitor)	$\Rightarrow$	$\theta_{Z_s} = 0$
(3)	$X_1 > 0$ (inductor),	$X_2 < 0$ (capacitor)	$\Rightarrow$	$\theta_{Z_s} = \pi$
(4)	$X_1 < 0$ (capacitor),	$X_2 > 0$ (inductor)	$\Rightarrow$	$\theta_{Z_s} = \pi$

Refer to Table 5.1 and note that cases 3 and 4 are impossible, since $\theta_{Z_s} = \pi$ implies that $Z_s = Z_3 + j(X_1 + X_2)$ has a negative real part. If Z_3 is a passive element, this is not possible. The remaining two possibilities yield two circuit configurations for oscillators of the type under consideration. They are shown in Figure 5.6. Notice that in each case Z_3 is allowed to have a positive real part which is represented by R.

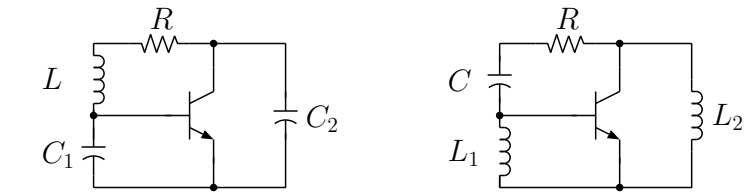


Figure 5.6: (a) $X_1 < 0$, $X_2 < 0$ $\Rightarrow$ Colpitts, (b) $X_1 > 0$, $X_2 > 0$ $\Rightarrow$ Hartley

The frequency of oscillation is easily found by finding the frequency where the phase angle of $Z_s = Z_3 + j(X_1 + X_2)$ is equal to zero. This will cause the phase angle of A_{lo} to be zero as well. The results are:

$$\begin{aligned} \text{Colpitts} \qquad \omega_o &= \frac{1}{\sqrt{L\frac{C_1C_2}{C_1+C_2}}} \\ \text{Hartley} \qquad \omega_o &= \frac{1}{\sqrt{C(L_1+L_2)}} \end{aligned} \tag{5.6}$$

These frequencies of oscillation are simply the resonant frequencies of the networks that result when the transistor is removed from the circuits shown in Figure 5.6. In fact, if Z_3 was lossless, then Z_1, Z_2 and Z_3 would constitute a lossless resonant circuit. A lossless circuit will oscillate indefinitely at the resonant frequency once it is initially excited. When Z_3 is lossy, then the oscillations cannot be maintained unless the transistor is added to make up for the losses.

So far we have considered only half of the Barkhausen criterion ($\arg[A_{lo}] = 0$). The results yielded the potential frequency of oscillation. For oscillations to occur, the gain of the transistor must be large enough to make A_{lo} equal to 1 at ω_o. The results of applying

this constraint are summarized below:

$$\text{Colpitts:} \quad |A_{lo}|_{\omega=\omega_o} = \frac{g_m}{R\omega_o^2 C_1 C_2} = 1 \tag{5.7}$$

$$\text{or,} \quad g_m = \frac{R(C_1+C_2)}{L}$$

$$\text{Hartley:} \quad |A_{lo}|_{\omega=\omega_o} = \frac{g_m \omega_o^2 L_1 L_2}{R} = 1 \tag{5.8}$$

$$\text{or,} \quad g_m = \frac{RC(L_1+L_2)}{L_1 L_2}$$

The values of g_m obtained in equations 5.7 and 5.8 are the values necessary for the circuit to support steady-state oscillations. We will denote these values by $g_{m,ss}$. In practical applications the transistor is biased to set the transconductance to a value somewhat larger, e.g. a factor of 2 to 5 larger, than $g_{m,ss}$. Setting $g_m > g_{m,ss}$ causes the loop gain at ω_o to be larger than 1 by the factor $g_m/g_{m,ss}$. One reason for doing this is to ensure that oscillations start reliably even if component values change slightly. Setting $A_{lo}|_{\omega=\omega_o} > 1$ means that the oscillations will not be maintained at a steady state; rather, they will grow in amplitude. Growth will proceed until the active device is no longer operating in the "small-signal" mode. As the oscillation grows, eventually the amplitude of the oscillation will be limited by nonlinear effects. To first order, the onset of nonlinear operation coincides with a reduction in the gain of the active device. This is called gain saturation, and the effect can be modeled as a decrease in the transconductance, g_m, and hence a decrease in the loop gain. As the amplitude of the oscillation grows, the transconductance is decreased to the point where the magnitude of the loop gain is 1. At this point steady-state oscillation will be maintained.

Note that previous considerations did not specify which of the transistor terminals was at RF ground. Thus they apply without modification to common base, common emitter or common collector circuits. Also, it was assumed that Z_1 and Z_2 were purely reactive. If this is not the case, one must go back to the general expression for loop gain:

$$A_{lo} = \frac{-g_m Z_1 Z_2}{Z_1 + Z_2 + Z_3} \tag{5.9}$$

Using this expression it is possible to find the frequency of oscillation and the minimum g_m required for oscillation.

5.3 Oscillator Analysis using Negative Resistance

The requirements for oscillation in a circuit can be derived using the so-called *negative resistance* concept. Consider the lossless LC circuit shown in Figure 5.7. The Kirchoff's loop equation for this circuit is $I(jX_C + jX_L) = 0$. Hence, finite current I is allowed if $X_L + X_C = 0$, which occurs at the resonant frequency, $\omega_o = \frac{1}{\sqrt{LC}}$.

Consider generalizing this idea to a circuit consisting of arbitrary impedances Z_A and Z_B, as shown in Figure 5.8. The Kirchoff's loop equation is $I(Z_A + Z_B) = 0$. Hence, finite

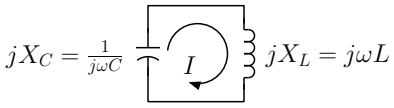


Figure 5.7: A lossless LC circuit supports finite current, I, if $X_L + X_C = 0$.

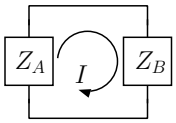


Figure 5.8: Finite current, I, is supported if $Z_1 + Z_2 = 0$.

current I is allowed if $Z_A + Z_B = 0$. Writing $Z_A = R_A + jX_A$ and $Z_B = R_B + jX_B$, it is apparent that the conditions for finite current are:

$$X_A + X_B = 0 \tag{5.10}$$

$$R_A + R_B = 0 \tag{5.11}$$

Equation 5.10 is called the *resonance condition*, and says that the reactances of Z_A and Z_B must sum to zero. Equation 5.11 specifies that the real parts of the impedances Z_A and Z_B must sum to zero. If one of the impedances has a positive real part, then the other impedance must have a negative real part, i.e. must exhibit *negative resistance*. This idea can be used to analyze and design oscillator circuits. The conditions represented by equations 5.10 and 5.11 are analogous to the two conditions for oscillation given by equations 5.2 and 5.3.

In practice the negative resistance concept is applied by breaking a candidate oscillator circuit into two parts, which are associated with the impedances Z_A and Z_B. Most oscillators consist of a single active device, so when the circuit is divided into two parts, one part will contain only passive components. This part will exhibit an impedance with positive real part. In order for oscillations to occur in the circuit, the part containing the active device must have an impedance with a negative real part.

It is instructive to analyze the oscillator circuit shown in Figure 5.3 using this approach. Suppose that the circuit is divided into two parts, by defining $Z_A = Z_1$. The impedance Z_B is the impedance seen looking into the port defined by the base and emitter terminals of the transistor, with Z_1 removed from the circuit. If the transistor is modeled using the simplified hybrid-pi model (Figure 5.4), the impedance between the base and emitter of the transistor is easily shown to be

$$Z_B = \frac{Z_2 + Z_3}{1 + g_m Z_2}.$$

Enforcing $Z_A + Z_B = 0$ yields

$$Z_1 + \frac{Z_2 + Z_3}{1 + g_m Z_2} = 0.$$

Multiplying both sides of this equation by $\frac{1+g_m Z_2}{Z_1+Z_2+z_3}$ yields the result

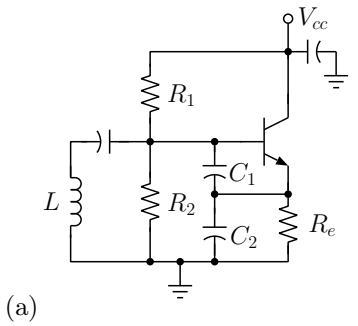
$$\frac{-g_m Z_1 Z_2}{Z_1 + Z_2 + Z_3} = 1,$$

which is the same as setting the loop gain, given in equation 5.4, to 1. Thus, the negative resistance analysis yields the same result as the loop gain analysis.

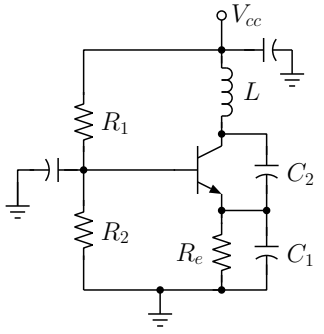
5.4 Example - Common-collector Colpitts Oscillator

5.4.1 Analysis

To illustrate some of the properties of oscillators, we will consider a realistic circuit in detail. Practical circuits for Colpitts-type oscillators implemented with the common-collector and common-base topologies are shown in Figure 5.9. The common-emitter variant is not shown. We will analyze the common-collector configuration shown in Figure 5.9a. The small-signal



(a)



(b)

Figure 5.9: (a) Common-collector Colpitts oscillator circuit, (b) Common-base Colpitts oscillator circuit.

equivalent circuit is shown in Figure 5.10, where the transistor has been replaced with its hybrid-pi model, and the finite Q of the inductor is modeled with a series resistance, r. The small-signal equivalent circuit can be further simplified by combining some elements as shown in Figure 5.11 , where

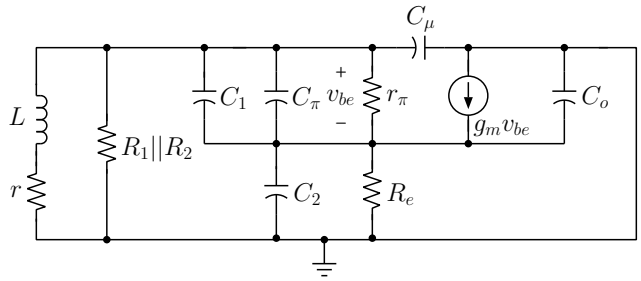


Figure 5.10: Small-signal equivalent circuit for common-collector Colpitts. The resistance, r, represents the series resistance of the inductor.

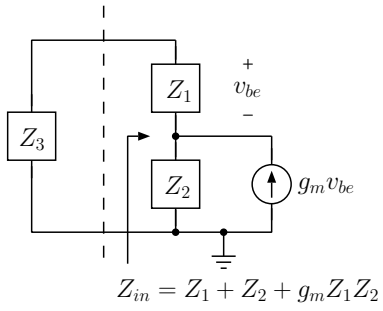


Figure 5.11: Simplified small-signal equivalent circuit of common-collector Colpitts oscillator.

$$\begin{aligned} Z_1 &= \frac{1}{j\omega(C_1+C_\pi)}||r_\pi \\ Z_2 &= \frac{1}{j\omega(C_2+C_o)}||R_e \\ Z_3 &= R_1||R_2||(r+j\omega L)||\frac{1}{j\omega C_\mu} \end{aligned} \tag{5.12}$$

In practice it is useful to choose $C_1 \gg C_\pi$, $C_2 \gg C_o$. This ensures that the external components swamp the internal capacitances of the transistor, thereby minimizing the circuit's dependence on variations in the internal transistor capacitances. Such variations could be caused by differences between individual transistors, or by variations in junction capacitances as a function of the voltage across the junctions. It is also useful to choose C_1 and C_2 to be large enough so that $1/(\omega C_1) \ll r_\pi$, $1/\omega C_2 \ll R_e$. This causes Z_1 and Z_2 to be dominated by the capacitances external to the transistor, thereby minimizing dependence on r_π (which depends on bias current and transistor β) and losses in r_π and R_e. In view of these considerations, in the following analysis we shall make the following replacements: $C_1 + C_\pi \to C_1$ and $C_2 + C_o \to C_2$. In other words, C_1 should be interpreted as the total capacitance across the base-emitter junction, including the internal capacitance of the transistor and any external capacitance. A similar interpretation holds for C_2.

It is now useful to make some approximations in order to simplify the analysis. We shall assume that the impedances of the capacitances C_1 and C_2 are small compared to r_π and R_e, respectively. In other words, define $Q_1 = \omega C_1 r_\pi$ and $Q_2 = \omega C_2 R_e$; we assume that $Q_1 \gg 1$ and $Q_2 \gg 1$. Then Z_1 and Z_2 can be transformed using a high-Q parallel to series transformation, i.e.

$$Z_1 \simeq \frac{r_\pi}{Q_1^2} + \frac{1}{j\omega C_1} = \frac{1}{\omega^2 C_1^2 r_\pi} + \frac{1}{j\omega C_1}, \quad Q_1 \gg 1 \tag{5.13}$$

$$Z_2 \simeq \frac{R_e}{Q_2^2} + \frac{1}{j\omega C_2} = \frac{1}{\omega^2 C_2^2 R_e} + \frac{1}{j\omega C_2}, \quad Q_2 \gg 1. \tag{5.14}$$

We shall also assume that the impedance of the Z_3 branch is well approximated by the impedance of the inductor, i.e.

$$Z_3 \simeq r + j\omega L. \tag{5.15}$$

If this is not the case, then r and L should be interpreted as the effective series resistance, and inductance, of Z_3.

The circuit has been split into two parts by the dashed line in Figure 5.11. Denote the impedance looking into the part of the circuit to the right of the dashed line by Z_{in}. Then

$$Z_{in} = Z_1 + Z_2 + g_m Z_1 Z_2.$$

The condition for steady-state oscillation is $Z_{in} + Z_3 = 0$, or,

$$Z_1 + Z_2 + Z_3 + g_m Z_1 Z_2 = 0. \tag{5.16}$$

The requirements for steady-state oscillation are obtained using equations 5.13-5.15 in 5.16. The real part of this equation is

$$\frac{1}{\omega^2 C_1^2 r_\pi} + \frac{1}{\omega^2 C_2^2 R_e} + r + g_m\left(\frac{1}{\omega^4 C_1^2 C_2^2 r_\pi R_e} - \frac{1}{\omega^2 C_1 C_2}\right) = 0. \tag{5.17}$$

It is now necessary to remember that r_π depends on g_m, i.e. $r_\pi = \beta/g_m$, so equation 5.17 is really quadratic in g_m, and the solution is somewhat messy. However, at sufficiently high frequencies, the term involving ω^{-4} may be neglected. More precisely, the term involving ω^{-4} may be neglected provided that $\omega^2 C_1 C_2 r_\pi R_e \gg 1$, or if $Q_1 Q_2 \gg 1$. Since we have already assumed that $Q_1 \gg 1$ and $Q_2 \gg 1$, neglect of this term is justified within the context of our original assumptions. Hence, neglecting the ω^{-4} term, and using $r_\pi = \beta/g_m$ in equation 5.17, the net resistance becomes

$$\frac{g_m}{\omega^2 C_1^2 \beta} + \frac{1}{\omega^2 C_2^2 R_e} + r - \frac{g_m}{\omega^2 C_1 C_2} = 0. \tag{5.18}$$

This equation can be solved for the transconductance needed to set the net resistance to zero at any particular frequency. This transconductance will be denoted $g_{m.ss}$ because it is the transconductance needed to support steady-state oscillations. The steady-state transconductance can be written as follows:

$$g_{m,ss} = \frac{\omega^2 C_1 C_2 r + \frac{C_1}{C_2 R_e}}{1 - \frac{C_2}{C_1 \beta}}. \tag{5.19}$$

Remember that this result was derived under the assumption that $Q_1 \gg 1$, and $Q_2 \gg 1$. The first assumption requires that $g_m \ll \omega C_1 \beta$, and should be checked after $g_{m,ss}$ is calculated from equation 5.19, however in most practical applications this inequality will be satisfied.

If R_e is allowed to approach infinity (so that Z_2 becomes a pure reactance), and if $\beta \gg \frac{C_2}{C_1}$, then $g_{m,ss}$ reduces to $g_{m,ss} = \omega^2 C_1 C_2 r$, the same as the result given in equation 5.7, which was derived by assuming that Z_1 and Z_2 were pure reactances. In the next section, it will be shown that it is desirable to make $C_1 > C_2$. Thus, in practical cases with $\beta \gg 1$, the second term in the denominator of 5.19 can be neglected, in which case

$$g_{m,ss} \simeq \omega^2 C_1 C_2 r + \frac{C_1}{C_2}\frac{1}{R_e}. \tag{5.20}$$

It is worth noting that equation 5.20 can be written in terms of the inductor Q_L. In terms of the inductor reactance, X_L, and the inductor Q_L (both are assumed to be known, or specified, at the frequency of oscillation) $r = X_L/Q_L$. As long as $Q_1 \gg 1$ and $Q_2 \gg 1$, the frequency of oscillation will be the frequency where the inductor reactance resonates with the series combination of C_1 and C_2, hence $X_L = \omega_o L \simeq \frac{1}{\omega_o \frac{C_1 C_2}{C_1 + C_2}}$. Thus, $r = X_L/Q_L \simeq (C_1 + C_2)/(\omega_o C_1 C_2 Q_L)$ - if this substitution is made in equation 5.20, the result is:

$$g_{m,ss} \simeq \frac{\omega_o (C_1 + C_2)}{Q_L} + \frac{C_1}{C_2}\frac{1}{R_e}. \tag{5.21}$$

In practical circuits additional reactances or resistances may be placed in parallel with the inductor. To implement a tunable oscillator it is common practice to use a variable capacitor in series or in parallel with the inductor. Since power must be extracted from the oscillator to drive other stages, a resistance representing an external load may be placed in shunt with the inductor if the output is taken across the inductor. The existing analysis is easily modified to handle such cases. In the original model, the resistance, r, represented the series resistance associated with the inductor. If the additional components are in parallel with the inductor, it is simply necessary to determine the series representation of the resulting Z_3 branch (at ω_o) and use the real part of this impedance in place of r in

equation 5.20. Alternatively, evaluate the component Q of the Z_3 branch at ω_o, and use this value in place of Q_L in equation 5.21. In some circuits the output is taken from the emitter - in this case, the external load resistance will appear in parallel with R_e, and the value used for R_e can be changed to include the contribution from the load.

As mentioned previously, the oscillator circuit will be designed such that the small-signal loop gain is larger than 1. The loop gain is equal to the ratio $g_m/g_{m,ss}$. As the oscillation amplitude builds up, the base-emitter voltage swing will increase. Transistor operation moves out of the small-signal linear regime, nonlinear effects become important, and the effective transconductance is reduced. Oscillation amplitude will grow until the effective large-signal transconductance is reduced to $g_{m,ss}$, which is the value required to sustain steady-state oscillations. The loop gain is the ratio of the initial transconductance to $g_{m,ss}$ ($A_{lo} = g_m/g_{m,ss}$), and is an important parameter that determines how deeply the transistor must be driven into the nonlinear regime before the steady-state condition is reached. In the next section we will see that under certain realistic assumptions, the ratio $g_m/g_{m,ss}$ determines the steady-state amplitude of the base-emitter voltage swing. This means that it is possible to predict, at least approximately, the amplitude of the voltage swing in the oscillating circuit.

5.4.2 Numerical Simulation

In order to test the results derived in the preceeding section, a common collector Colpitts circuit was simulated using the Agilent ADS simulation program. The frequency of oscillation was chosen to be $f_o = 50\,\text{MHz}$. The circuit was simulated using a 2N5179 transistor model. To compare the simulated results with analytical predictions, the parameters of the hybrid-pi model are needed. The transistor model was probed using simulations to determine that $\beta = 62$, $C_\pi \simeq 10\,\text{pF}$, $C_\mu \simeq 1\,\text{pF}$. The capacitance C_o was assumed to be zero. A supply voltage of 12.0V was used. The DC blocking capacitor (in series with the inductor) was given a value of $0.01\,\mu\text{F}$ for the simulation. The values chosen for the inductor and bias resistances are:

$$\begin{aligned} L &= 500\,\text{nH} \\ Q_L &= 50 \\ R_1 &= 50\,\text{k}\Omega \\ R_2 &= 50\,\text{k}\Omega \\ R_E &= 5\,\text{k}\Omega \end{aligned}$$

The bias network is designed to give a quiescent collector current of approximately 1 mA. Thus $r_\pi \simeq 1.55\,\text{k}\Omega$ and $g_m \simeq 40\,\text{mS}$. The inductor has $Q_L = 50$, however the bias resistors R_1 and R_2 appear in parallel across the inductor. This lowers the effective Q of the inductor to Q_L=38, which is the value that was used in the analytical calculations. [1] The effect of C_μ, which also appears across the inductor, is small and was ignored. To start the oscillations, the circuit was artificially excited by pulsing the supply voltage from 12.0 V to 12.1 V and back to 12.0 V within a period of 10 ns. This is necessary in a numerical simulation in order to provide some initial disturbance that can then grow into a steady-state oscillation. In practice, omnipresent thermal noise (or the step excitation provided by connecting the supply voltage) would play this role.

[1] A common base configuration would be better in this respect, as the bias resistors do not load the inductor in that circuit.

Six different $\{C_1, C_2\}$ pairs were considered (see Table 5.2). For each case, the transconductance required to sustain steady-state oscillation ($g_{m,ss}$), and the small-signal loop gain ($g_m/g_{m,ss}$) are given. The values of C_1 and C_2 were chosen to satisfy $2\pi(50\,\text{MHz}) = \sqrt{L\frac{C_1C_2}{C_1+C_2}}^{-1}$, so that the oscillation frequency is fixed at approximately 50 MHz in all cases. Notice that if a $\{C_1, C_2\}$ pair is found that resonates with L at the desired ω_o, the values can be reversed without significantly changing the resonant frequency. Compare cases 4 and 6 for an example. As we shall see, however, the case with $C_1 > C_2$ will usually give better performance.

Table 5.2: Parameters used for oscillator simulations.

Case	C_1	C_2	$g_{m,ss}$	$g_m/g_{m,ss}$
1	2500 pF	20.4 pF	45.3 mS	0.88
2	1000 pF	20.6 pF	18.1 mS	2.2
3	500 pF	21.1 pF	9.42 mS	4.4
4	200 pF	22.5 pF	3.77 mS	11.0
5	100 pF	25.4 pF	1.83 mS	21.8
6	22.5 pF	200 pF	2.23 mS	17.9

Simulation results are summarized in Figures 5.12-5.17. In each Figure, the upper left plot shows the voltage across the inductor for an interval of 1 μs after the initial transient. The other plots show the emitter current, base-emitter voltage, and emitter voltage for the period 0.9-1.0 μs after the initial transient. For the cases where oscillation occurs, the expanded plots show the current and voltage waveforms when the circuit is undergoing steady-state oscillation.

The loop gain ($g_m/g_{m,ss}$, column 4 in Table 5.2) is smaller than 1 for Case 1. Therefore the initial disturbance excites damped oscillations at a frequency approximately equal to f_o. This case illustrates the fact that sustained oscillations cannot develop if the small-signal loop gain is smaller than 1.

In cases 2 through 6 the loop gain is larger than 1, and the initial transient excites growing oscillations, as shown in the plot of the voltage across the inductor (upper left plot in each Figure). As the loop gain increases from 2.2 to 21.8 in cases 2 through 5 the initial transient builds up to steady-state condition faster, and the amplitude of the steady-state voltage swing increases. Notice that increases in the base-emitter voltage swing result in the collector current (lower, left) waveform exhibiting increasingly narrow and increasingly large current "spikes". For the larger loop gains, the transistor is essentially cut off for most of the oscillation cycle. The transistor injects a short current pulse into the resonator once each cycle, near the positive peak of the voltage swing. Thus, for most of the oscillation period, the circuit is un-excited and oscillation is maintained by the "flywheel effect" of the high-Q resonant circuit composed of L, C_1, and C_2. The emitter current flows into an impedance that has large magnitude at ω_o, and very small magnitude at harmonics of ω_o - hence, the emitter voltage is nearly sinusoidal. In cases 4 and 5 where the loop gain is largest, and the current spikes are narrowest, some distortion of the emitter voltage waveform can be seen near the voltage peaks. In general, larger loop gains are associated with more distortion in the output waveform.

Case 6 was chosen to illustrate what happens when $C_1 < C_2$. Notice that the loop gain

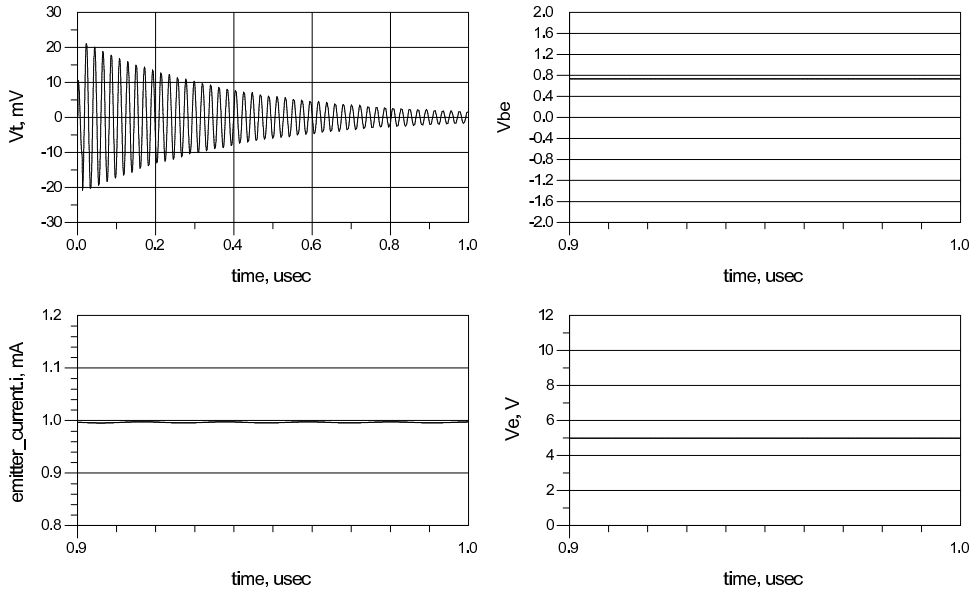


Figure 5.12: Case 1 - loop gain is 0.88, which is less than one, so oscillation does not start.

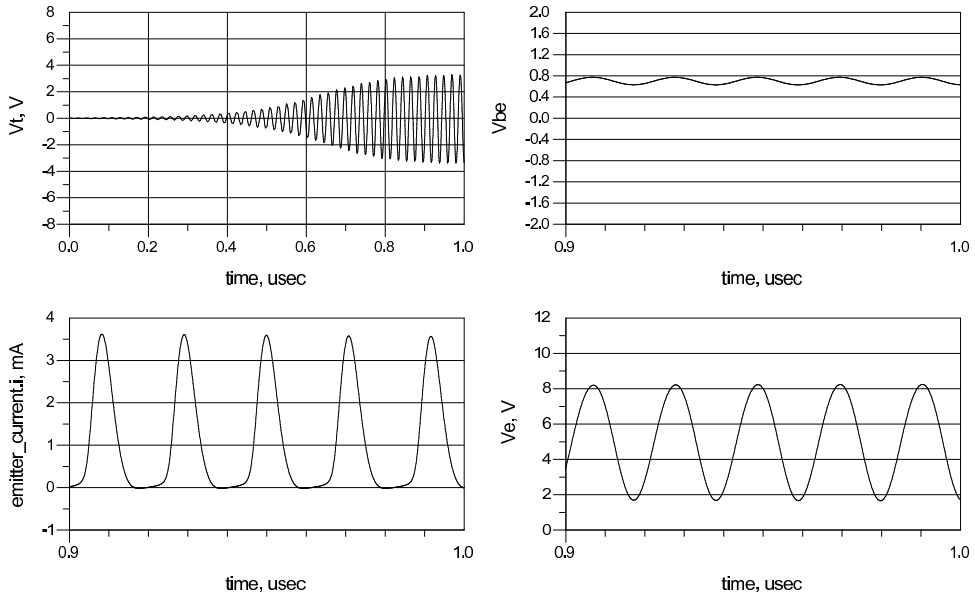


Figure 5.13: Case 2 - $g_m/g_{m,ss} = 2.2$.

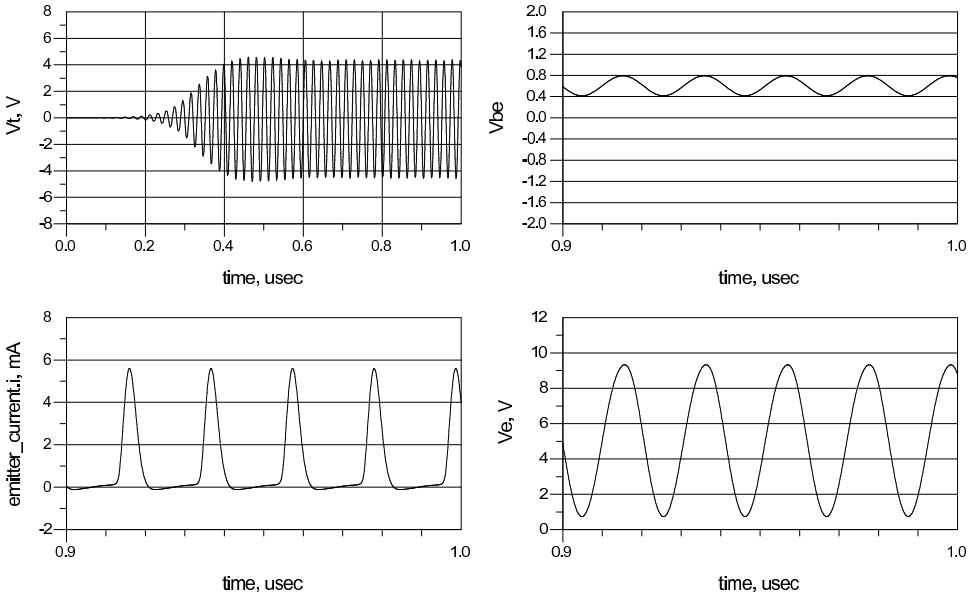


Figure 5.14: Case 3 - $g_m/g_{m,ss} = 4.4$.

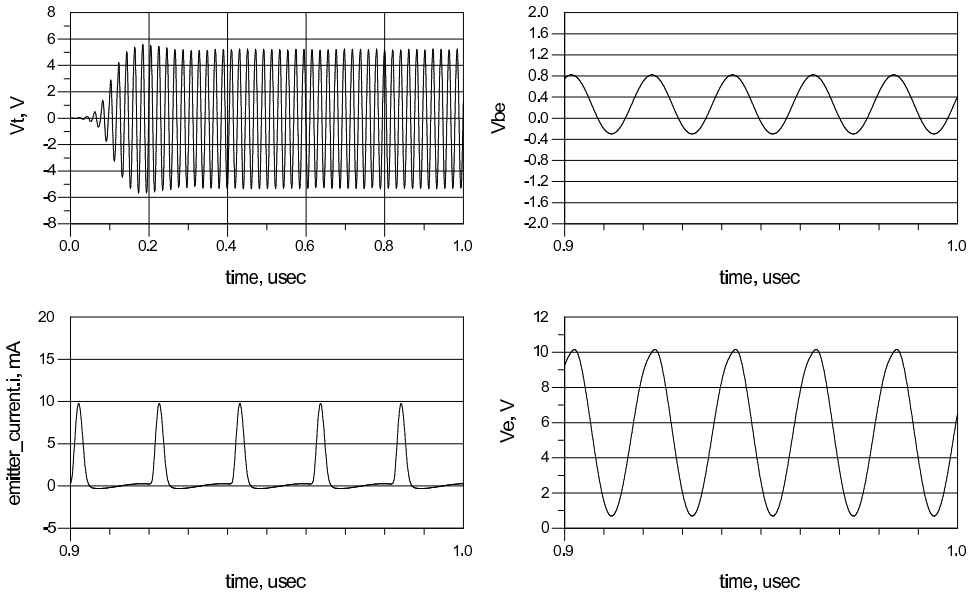


Figure 5.15: Case 4 - $g_m/g_{m,ss} = 11.0$.

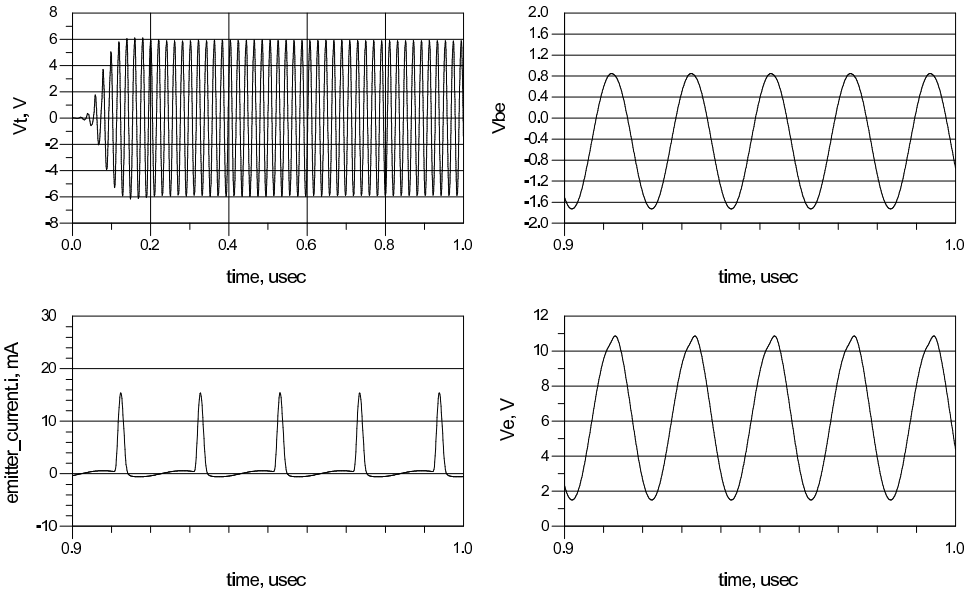


Figure 5.16: Case 5 - $g_m/g_{m,ss} = 21.8$.

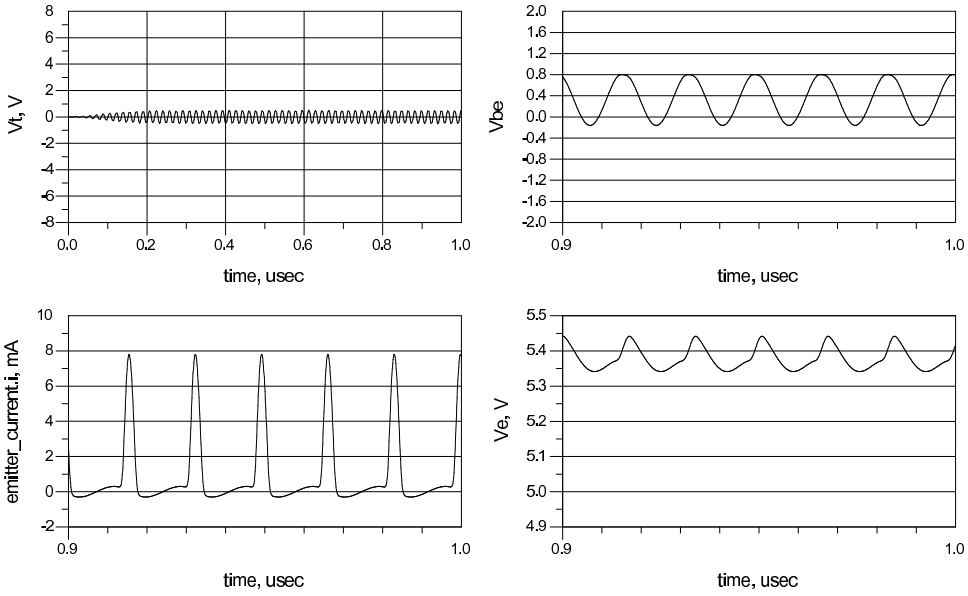


Figure 5.17: Case 6 - $g_m/g_{m,ss} = 17.9$.

in case 6 is not very different from the loop gain in case 5 (17.9 vs. 21.8), yet the voltage swing across the inductor and the emitter voltage swing are quite small. Furthermore, the emitter voltage is clearly non-sinusoidal, indicative of relatively high harmonic content. This reflects the fact that C_1 is not large enough to swamp r_π, and the resulting dissipation in r_π results in a lower overall Q for the LC tank circuit than in case 4. As a result, the harmonics contained in the spiky emitter current waveform are more prominent in the emitter voltage waveform. The output from the oscillator would be taken from the top of the inductor, or from the emitter - so comparison of cases 4 and 6 shows that choosing $C_1 > C_2$ (as in case 4) yields a larger output swing, and less distortion in the output.

Cases 4 and 6, which have comparable loop gains but very different amplitudes of oscillation, motivate the need to understand the factors which govern the amplitude of oscillations. Consider the large-signal equivalent circuit for the oscillator circuit in Figure 5.18.

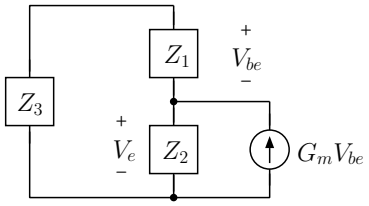


Figure 5.18: A model for the Colpitts oscillator when oscillating in steady-state. The small-signal transconductance has been replaced by a large signal transconductance, G_m, which is a function of the base-emitter voltage swing, $|V_{be}|$.

The small-signal transconductance has been replaced by an effective (large-signal) transconductance (G_m) in Figure 5.18. If the circuit is undergoing steady-state oscillations, the large signal transconductance will be equal to the transconductance required to sustain oscillations, i.e. $G_m = g_{m,ss}$. In Appendix A, it is shown that when a BJT is biased for constant DC emitter current, and when the base-emitter voltage swing is sinusoidal at frequency ω_o, a large-signal transconductance can be defined, which relates the amplitude of the emitter current component at ω_o to the amplitude of V_{be}. The large signal transconductance can be written as

$$G_m(x) = g_m \frac{2}{x} \frac{I_1(x)}{I_0(x)}, \tag{5.22}$$

where g_m is the small-signal transconductance, and $x = |V_{be}|/(25\,\text{mV})$ is the normalized, peak base-emitter voltage amplitude. Figure 5.19 shows how the ratio G_m/g_m decreases as $|V_{be}|$ increases. In steady-state, $|V_{be}|$ settles at the value required to set the large signal transconductance equal to the minimum value required to sustain oscillations, i.e. in steady state $G_m(|V_{be}|/(25\,\text{mV})) = g_{m,ss}$. For example, consider case 2, where the small-signal loop gain is $g_m/g_{m,ss} = 2.2$. To reach the steady state, $|V_{be}|$ must increase until $G_m/g_m = 1/2.2 = 0.45$. From Figure 5.19 it is apparent that this occurs when $|V_{be}|/(25\,\text{mV})$ is approximately equal to 3.8. A more exact value can be obtained using equation 5.22; numerically solving the equation $0.45 = \frac{2}{x} I_1(x)/I_o(x)$ yields $x \simeq 3.75$. The predicted steady-state amplitude of $|V_{be}|$ is then $(3.75)(25\,\text{mV}) = 94\,\text{mV}$, which compares favorably with what is observed in the simulations. Table 5.3 compares the base-emitter swings observed in the simulations with values predicted in this way. In cases 2-6, where oscillation occurs, the

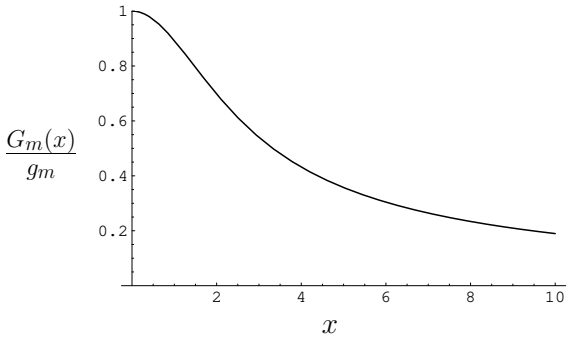


Figure 5.19: Ratio of large signal to small-signal transconductance. The parameter $x = \frac{|V_{be}|}{25\,\mathrm{mV}}$.

Table 5.3: Comparison between predicted and simulated $|V_{be}|$ and $|V_e|/|V_{be}|$ ratio.

Case	$g_m/g_{m,ss}$	$\lvert V_{be}\rvert$ pred.	$\lvert V_{be}\rvert$ sim.	$\lvert V_e\rvert$ sim.	C_1/C_2	$\lvert V_e\rvert/\lvert V_{be}\rvert$
1	0.88	-	-	-	122.5	-
2	2.2	94 mV	71 mV	3.27 V	48.5	46.1
3	4.4	207 mV	185 mV	4.30 V	23.7	23.2
4	11.0	537 mV	554 mV	4.70 V	8.9	8.5
5	21.8	1.08 V	1.29 V	5.40 V	3.9	4.2
6	17.9	882 mV	483 mV	*	0.11	*

prediction yields a reasonable approximation to the simulated values.

Once $|V_{be}|$ is known, or has been predicted, the amplitude of the emitter voltage, $|V_e|$, can be estimated. A node equation at the junction between the current source, Z_1, and Z_2 in Figure 5.22 yields

$$\frac{V_e}{V_{be}} = \frac{Z_2}{Z_1} + Z_2 g_{m,ss} \tag{5.23}$$

$$= \frac{C_1}{C_2} + Z_2 g_{m,ss}. \tag{5.24}$$

The second term will be small compared to the first in practical cases. This can be verified for the cases considered here by inserting numerical values, or by noting that the simulation results show that V_e and V_{be} are nearly in-phase. Hence, the ratio V_e/V_{be} is real - confirming that the second term is negligible in all cases considered here. Thus, the emitter voltage swing is determined by the capacitive transformer consisting of C_1 and C_2 and can be written as:

$$|V_e| = |V_{be}|\frac{C_1}{C_2}. \tag{5.25}$$

Table 5.3 provides data showing how the ratio $|V_e|/|V_{be}|$ determined from simulation compares to the C_1/C_2 ratio. In cases 2-5 the agreement is excellent. In case 6, the emitter voltage waveform was non-sinusoidal, and the amplitude of the fundamental component is not easily extracted from the time-domain waveform, so the ratio was not calculated. These results show that making the ratio $C_1/C_2 > 1$ will cause the emitter (output) voltage to be larger than the base-emitter voltage swing. When $C_1/C_2 > 1$, increasing C_1/C_2 yields smaller loop gains, so oscillators with large C_1/C_2 tend to have more sinusoidal output waveforms because the base-emitter swing is relatively small, and the emitter current is more sinusoidal.

5.5 Example: Voltage Controlled Oscillator (VCO)

Figure 5.20 is an example of an oscillator circuit used in a commercial television receiver. The oscillator is part of the VHF tuner that performs the function of converting the VHF channels to the IF frequency near 45 MHz, this local oscillator is voltage-tuned using a varactor diode. Figure 5.20 also illustrates how different inductances can be switched in and out of the circuit using diode switches. Although this circuit looks more complicated than the one we have been studying, it is essentially the same, except for the fact that the base of the transistor is at RF ground (common-base configuration).

A simplified schematic of Figure 5.20 is shown in Figure 5.21.

Based on earlier analysis, it should be clear the the approximate frequency of oscillation for this circuit will be the frequency where C_1, C_2 and the L-C_v combination are resonant. We also know that the frequency of oscillation will be such that the L-C_v combination looks inductive.

Thus ω_o is the solution to

$$\frac{1}{j\omega_o C_S} + \frac{j\omega_o L \frac{1}{j\omega_o C_v}}{j\omega_o L + \frac{1}{j\omega_o C_v}} = 0 \tag{5.26}$$

where

$$C_S = \frac{C_1 C_2}{C_1 + C_2} \tag{5.27}$$

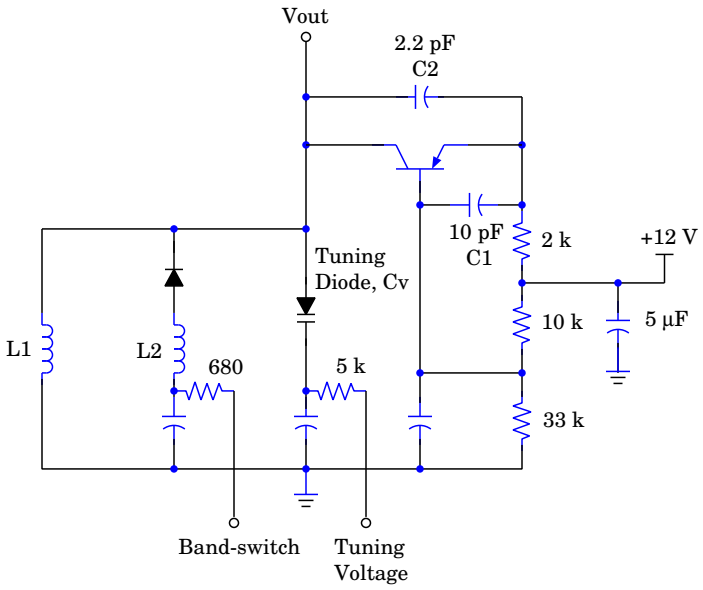


Figure 5.20: Example: VHF oscillator for TV tuner.

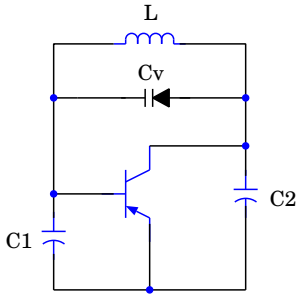


Figure 5.21: Simplified schematic of VHF oscillator circuit.

Therefore

$$\omega_o = \frac{1}{\sqrt{L(C_S + C_v)}} \tag{5.28}$$

Equation 5.28 can be used to find the range of values over which C_v must vary in order to cover a certain range of oscillation frequencies.

Voltage controlled oscillator's (VCO's) find extensive application in phase-locked loops (PLL's) which in turn are used for a majority of frequency synthesizer applications. For PLL or synthesizer design, an important parameter is the so-called VCO "gain constant," which is simply the incremental change in oscillation frequency for an incremental change in varactor bias voltage. Typically, the varactor would be an "abrupt" transition type, so that

$$C_v = \frac{C^{(1)}}{\sqrt{V_D}}$$

where V_D is the varactor bias voltage and $C^{(1)}$ is the junction capacitance at 1 volt of bias. The gain constant is

$$K_o = \frac{\partial \omega_o}{\partial V_D}\Big|_{\omega_o = \text{oscillation frequency}} \tag{5.29}$$

$$= \frac{\partial \omega_o}{\partial C_v}\frac{\partial C_v}{\partial V_D}\Big|_{\omega_o = \text{oscillation frequency}}$$

Generally, K_o will be a nonlinear function of V_D. This implies that the gain constant will differ at various varactor bias settings, a factor which must be considered in the design of a PLL.

5.6 Oscillator Phase Noise

The most significant departure from ideal behavior in oscillators is the presence of phase noise, represented by a random variation in the oscillator's phase angle, i.e.

$$V(t) = V_o \cos[2\pi f_o t + \phi(t)] \tag{5.30}$$

where $\phi(t)$ is a random noise process that has various physical processes. Amplitude noise may also be present, but it is easy to remove this noise component by passing the signal through a limiter. Figure 5.22 shows measurements of the output spectrum of a synthesized signal generator operating at 10.240 MHz. The width of the narrow central spike is determined by the resolution bandwidth of the spectrum analyzer which was set to 100 Hz for this measurement. Residual phase modulation — *phase noise* — on the signal contributes the broad bandwidth pedestal underneath the spike. Phase noise spectral density is specified as a function of frequency separation from the carrier in terms of the noise power level within a 1 Hz bandwidth. The noise power level is given in dB referenced to the carrier power level. For example, in Figure 5.22 the triangular marker has been placed 1000 Hz above the carrier frequency, and the display indicates that the noise pedestal is at -74.2 dBc, where dBc means *relative to the carrier*. Assuming that the noise spectral density is constant within the resolution bandwidth the measured noise power level in 100 Hz bandwidth can

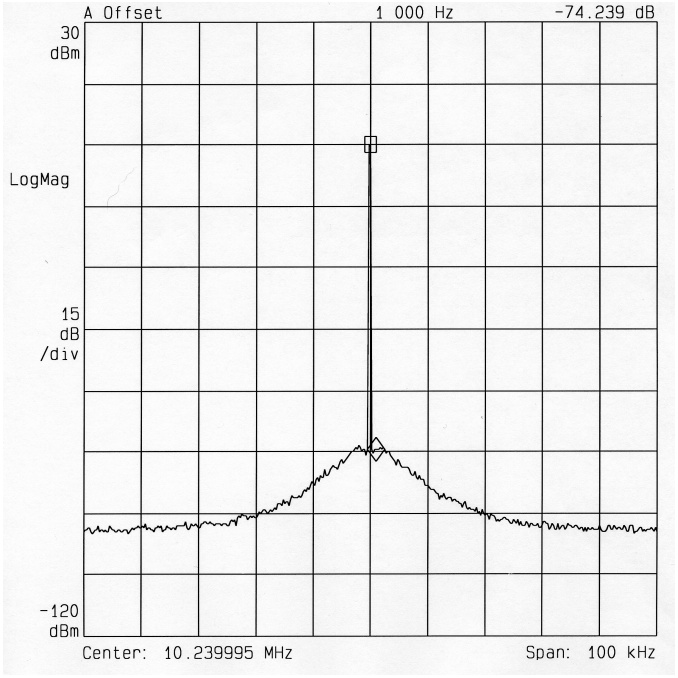


Figure 5.22: Spectrum analyzer display showing phase noise on the output signal from an HP E4432B synthesized signal generator. The signal generator's output frequency was 10.240 MHz and the output level was 0 dBm. The triangular marker is offset 1000 Hz from the carrier frequency, and the phase noise at this offset is -94.2 dBc/Hz (see text for explanation).

be scaled to 1 Hz bandwidth by dividing by 100 (or by subtracting $10\log_{10}(100) = 20$ dB). Thus, the phase noise level at 1000 Hz offset would be reported as $-74.2 - 20 = -94.2$ dBc/Hz.

Assuming that the random phase process is differentiable, the instantaneous frequency of the noisy oscillator can be written as

$$f_{inst} = f_o + \frac{1}{2\pi}\frac{d\phi(t)}{dt}. \tag{5.31}$$

The instantaneous frequency deviation is therefore

$$\delta f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt}. \tag{5.32}$$

The fractional deviation in instantaneous frequency can be written:

$$y(t) = \frac{\delta f(t)}{f_o} = \frac{1}{2\pi f_o}\frac{d\phi(t)}{dt}. \tag{5.33}$$

Measures of phase (or frequency) stability may be defined in the frequency domain in terms of a power spectral density, or in the time domain in terms of a phase, or frequency, or fractional-frequency variance defined over some time-interval.

If the random phase process has spectral density $S_\phi(f)$, then the corresponding spectral density for fractional frequency fluctuations is:

$$S_y(f) = \frac{f^2}{f_o^2}S_\phi(f). \tag{5.34}$$

The spectral density $S_y(f)$ will be invariant to multiplication or division of the frequency using ideal frequency multipliers or dividers because frequency multiplication or division scales the reference frequency, f_o, and the frequency deviation, δf, by the same factor.

In the time-domain, oscillator stability is usually characterized by the expected value of the sample variance of fractional frequency fluctuations. Define the average fractional frequency deviation over an interval τ, beginning at time t_k by

$$\begin{aligned} \overline{y}_k &= \tfrac{1}{\tau}\textstyle\int_{t_k}^{t_k+\tau} y(t)dt \\ &= \tfrac{\phi(t_k+\tau)-\phi(t_k)}{2\pi f_o \tau}. \end{aligned} \tag{5.35}$$

When N measurements of $\overline{y}_k$ are available, each taken over interval τ and with starting times separated by interval T, the sample variance of the measurements is:

$$\sigma_y^2(N,T,\tau) = \frac{1}{N-1}\sum_{n=1}^{N}(\overline{y}_n - \frac{1}{N}\sum_{k=1}^{N}\overline{y}_k)^2 \tag{5.36}$$

The expected value of the quantity σ_y^2 is known as the "Allan variance" of the oscillator. For most purposes, oscillator stability is specified in terms of the Allan variance for N=2 and $T = \tau$, i.e. the quantity that is specified most often is the variance of the difference between 2 successive measurements with no dead time between the measurement intervals. In this case, the Allan variance reduces to:

$$< \sigma_y^2(\tau) >= \frac{< (\overline{y}_2 - \overline{y}_1)^2 >}{2} \tag{5.37}$$

Thus, the Allan variance is the variance of the difference between two successive measurements of the fractional frequency deviation, measured over a time interval τ. The square-root of the Allan variance is called the "Allan Deviation", and is commonly supplied on oscillator data sheets where it is often plotted verses the measurement interval τ.

5.7 References

1. Clarke, Kenneth K. and Donald T. Hess, *Communication Circuits: Analysis and Design*, Addison-Wesley, 1978.

2. Frerking, Marvin E., *Crystal Oscillator Design and Temperature Compensation*, Van Nostrand Reinhold, New York, 1978.

3. Krauss, H. L., C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, John Wiley & Sons, New York, 1980.

4. Matthys, Robert J., *Crystal Oscillator Circuits*, John Wiley & Sons, New York, 1984.

5. Parzen, Benjamin, *Design of Crystal and other Harmonic Oscillators*, John Wiley & Sons, New York, 1983.

6. Smith, Jack, *Modern Communications Circuits*, Second Edition, McGraw Hill, 1998.

5.8 Homework Problems

1. Consider the voltage amplifier in Figure 5.23, with
   V_{cc}=12 V $\quad R_1 = 10\,\text{k}\Omega \quad R_2 = 30\,\text{k}\Omega \quad R_e = 1\,\text{k}\Omega$
   $R_C = \infty \quad R_L = 1\,\text{k}\Omega \quad L = 2\,\mu\text{H} \quad C = 50\,\text{pF}$
   Capacitors that are not labeled are assumed to be "short circuits" over the frequency range of interest. The transistor's β is large enough so that the bias point does not explicitly depend on its value. You may neglect the transistor parameters r_x, r_μ, C_μ, r_o, and C_o in your analysis for parts 1b, 1c, and 1d.

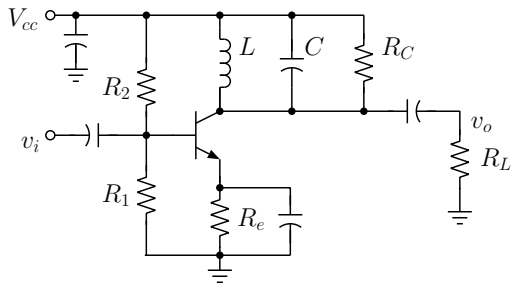


Figure 5.23: Common-emitter amplifier with tuned output.

   (a) Find the quiescent collector current, I_{CQ}. Express your result in mA.
   (b) Find the resonant frequency of the amplifier. The voltage gain will be largest at this frequency. Express your result in MHz. Do not make any "2π" errors!
   (c) Find the voltage gain at resonance.
   (d) Find the 3 dB bandwidth of the amplifier. Express your result in MHz.

2. The hybrid-pi parameters for a transistor are:

$$\begin{aligned} r_\pi &= \frac{0.025\,\beta}{I_{CQ}} \\ C_\pi &= 25\,\text{pF} \\ C_\mu &= 8\,\text{pF} \\ r_o &= 100\,\text{k}\Omega \\ \beta &= 100 \end{aligned} \tag{5.38}$$

   (a) The transistor, with parameters given above, is used in a common emitter amplifier, as shown in Figure 5.24. Find an expression for the voltage gain of amplifier. Express your result in terms of R_C, R_L, C_μ, and g_m. You may assume that r_o is much larger than $R_C \| R_L$ and can be neglected. Assume that r_x and C_o are small and may also be neglected.
   (b) If the quiescent collector current is 1mA, $R_C = R_L = 5k\,\Omega$, and all coupling and bypass capacitors are assumed to be perfect short circuits, sketch the magnitude and phase of the voltage gain as a function of frequency. What are the magnitude and phase of the transfer function at 15 MHz?

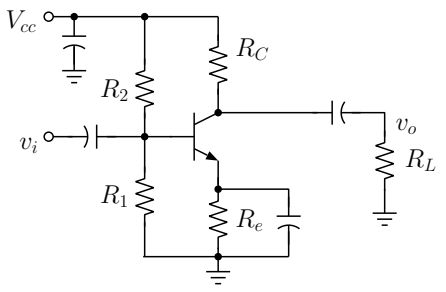


Figure 5.24: Common emitter amplifier

(c) Suppose that we attempt to measure the voltage gain of this amplifier using a 10X scope probe. The shunt capacitance of this type of probe is 12pF. Compute the gain (at 15 MHz) that would be measured. Compare to the result from 2b.

3. Consider the voltage amplifier shown in Figure 5.23. Capacitors that are not labeled are assumed to be short circuits over the frequency range of interest. The other circuit parameters are
V_{CC}=12 V $R_1 = 5\,\text{k}\Omega$ $R_2 = 10\,\text{k}\Omega$ $R_E = 3.3\,\text{k}\Omega$
$R_C = 5\,\text{k}\Omega$ $R_L = 1\,\text{k}\Omega$ $L = 20\,\mu\text{H}$ $C = 90\,\text{pF}$
Use the simplified hybrid-pi model shown in Figure 5.25 for the transistor:

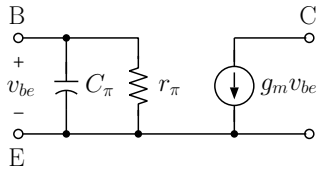


Figure 5.25: Simplified hybrid-pi model

(a) Find the resonant frequency of the amplifier. The voltage gain (v_o/v_i) will be largest at this frequency. Express your result in MHz. Do not make any "2π" errors!

(b) Find the 3 dB bandwidth of the amplifier. Express your result in MHz.

(c) Find the quiescent collector current, I_{CQ}. State any approximations.

(d) Find the voltage gain (v_o/v_i) at resonance.

4. A quartz crystal resonator has the equivalent circuit shown in Figure 5.26. Suppose it is known that a particular crystal has Q=50,000, series resonant frequency of 5 MHz, and parallel resonant frequency of 5.005 MHz. It is also known that the parallel resonant frequency shifts by 1 kHz, if a 3 pF capacitor is placed in parallel with the

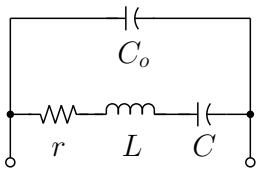


Figure 5.26: Equivalent circuit for a quartz crystal resonator.

crystal. Find the values of the equivalent circuit elements for the crystal. You may use whatever approximations are appropriate for a low-loss crystal.

5. Consider the AC equivalent circuit for an oscillator circuit in Figure 5.27:

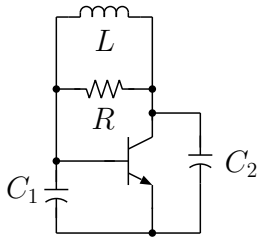


Figure 5.27: AC equivalent circuit for oscillator circuit

The loop gain for this circuit can be shown to be

$$A_{lo} = \frac{g_m \frac{1}{\omega^2 C_1 C_2}}{\frac{1}{j\omega C_1} + \frac{1}{j\omega C_2} + \frac{j\omega RL}{R + j\omega L}} \tag{5.39}$$

It has been assumed that the transistor immittances could be ignored in deriving the above equation. Find an expression for the frequency of oscillation, ω_o. To simplify interpretation of your result, define $C' = C_1 C_2/(C_1 + C_2)$ and express your result in terms of R, L, and C'.

6. Imagine that we could produce an active device, as shown in Figure 5.28, that has an impedance Z = -600 Ω. Assume that the device has this impedance at all frequencies. Suppose our idealized circuit is connected to an external R, L, C network as shown in the Figure and with R = 120 Ω, L = 4.0μH, and C = 52pF.

   (a) Will the circuit oscillate?

   (b) If so, find the frequency at which the circuit will oscillate. If not, find the frequency at which the circuit would oscillate if the negative resistance of the active source was decreased enough to cause oscillations to start.

7. We want to make an oscillator by cascading a number of identical stages as shown in Figure 5.29. Assume that the transistors can be represented by the simplified equivalent circuit in Figure 5.30 and that the capacitances C have finite values.

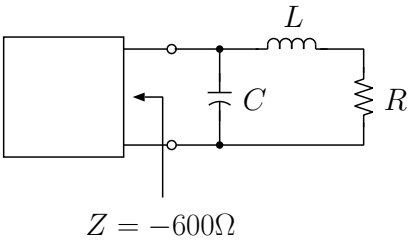


Figure 5.28: Idealized circuit

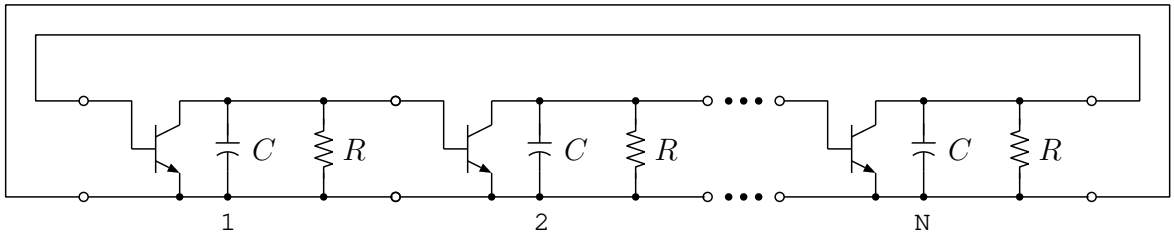


Figure 5.29: An oscillator consisting of N identical sections in cascade.

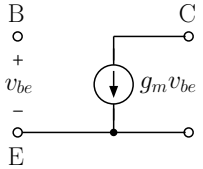


Figure 5.30: Simplified equivalent circuit

(a) Find the minimum number of stages (N) that are necessary in order for oscillation to occur.

(b) Assuming that we use the minimum number of stages, find an expression for the frequency of oscillation.

(c) Find the minimum value of g_m which will ensure that oscillation occurs.

8. Consider the circuit in Figure 5.31.
Use the simplified equivalent circuit for the transistor given in problem 7.

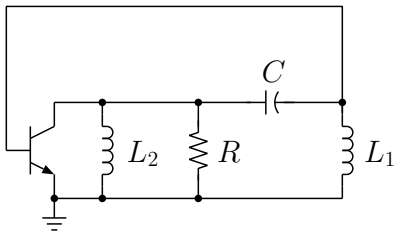


Figure 5.31: Hartley oscillator.

(a) Find an expression for the frequency of oscillation.

(b) Find the minimum value of g_m required to make the circuit oscillate.

9. Consider the oscillator circuit in Figure 5.32. The amplifier has infinite input impedance

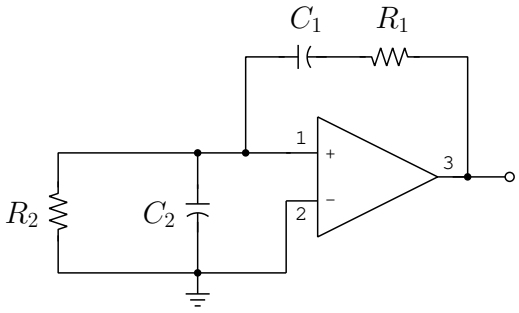


Figure 5.32: Wein Bridge oscillator circuit.

and voltage gain A. The equivalent circuit for the amplifier is shown in Figure 5.33. The numeric labels on the terminals correspond to the labels in Figure 5.32.

(a) Find an expression for the frequency of oscillation (if it occurs).

(b) Find an expression for the value of the voltage gain (A) required to sustain steady-state oscillations.

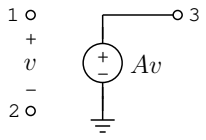


Figure 5.33: Equivalent circuit for amplifier

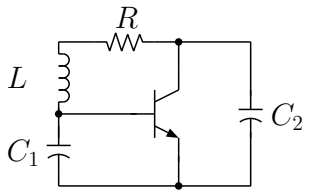


Figure 5.34: Colpitts oscillator

10. Consider the Colpitts oscillator in Figure 5.34 where R represents the series resistance of the inductor.

    The loop gain is given by Equation 5.40, if the transistor immittances are ignored:

$$A_{lo} = \frac{g_m/\omega^2 C_1 C_2}{R + j\omega L + \frac{1}{j\omega}(\frac{1}{C_1} + \frac{1}{C_2})} \tag{5.40}$$

    The frequency of oscillation is given by

$$\omega_o = \frac{1}{\sqrt{L\frac{C_1 C_2}{C_1 + C_2}}} \tag{5.41}$$

    A measure of the relative stability of the oscillator is the slope of the loop-gain phase characteristic evaluated at the oscillation frequency, i.e.,

$$\frac{d\phi}{d\omega}|_{\omega=\omega_o} \tag{5.42}$$

    where ϕ is the phase of the loop gain.

    (a) Show that $\frac{d\phi}{d\omega}|_{\omega=\omega_o} = \frac{-2Q}{\omega_o}$ where Q is the inductor Q evaluated at the frequency of oscillation.

    (b) Suppose that an oscillator is built at 5MHz with an inductor having a $Q = 50$. Estimate the shift in oscillation frequency, if the overall phase of the loop gain is perturbed by 5 degrees due to temperature drift of the capacitors.

    (c) Now assume that the inductor is replaced with a crystal having an effective $Q = 20,000$. Estimate the frequency shift.

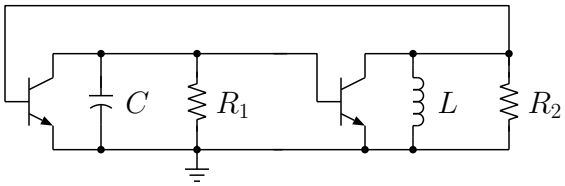


Figure 5.35: Two-stage oscillator circuit.

11. Consider the oscillator circuit in Figure 5.35.

    Assume that the transistors are identical and can be represented by the simplified equivalent circuit in Figure 5.36.

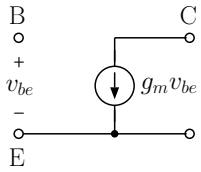


Figure 5.36: Simplified equivalent circuit.

    (a) Find an expression for the (non-zero) frequency of oscillation (if it occurs).

    (b) Find the minimum value of g_m required for oscillation to occur.

12. Consider the oscillator circuit in Figure 5.37 where a voltage amplifier with 50 Ω input and output impedance and open-circuit voltage gain A is used with a lossy section of transmission line having characteristic impedance $Z_o = 50\,\Omega$, length L, and attenuation and phase shift per unit length denoted by α and β, respectively. Since the line is terminated in its characteristic impedance, the voltage at the output of the line will be related to the voltage at the input of the line by the following equation:

$$V_{out} = V_{in} e^{-(\alpha + j\beta)L} \tag{5.43}$$

    where

$$\beta = \frac{2\pi f}{v_p} \tag{5.44}$$

    and v_p is the phase velocity for propagation on the line.

    (a) Find an expression for the potential frequencies of oscillation. Assume that the voltage gain, A, is real (zero phase shift) and that the input and/or output of the voltage amplifier is AC-coupled with a coupling capacitor (not shown) that will prevent the circuit from oscillating at zero frequency. The capacitor may be assumed to have no effect on the circuit performance at finite frequencies.

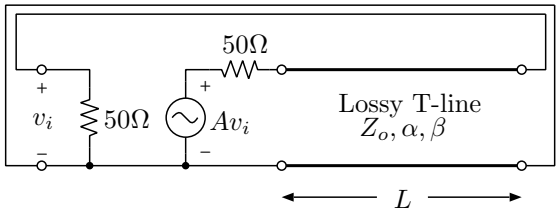


Figure 5.37: Oscillator circuit.

(b) Suppose that the T-line loss per unit length, α, is proportional to frequency, i.e., $\alpha = Kf$. Define the minimum and maximum values for the voltage gain, A, such that the circuit will oscillate only at the smallest (non-zero) frequency found in part 12a.

13. Suppose a capacitor is placed in parallel with a quartz crystal.

(a) Describe how the series and parallel resonant frequencies of the crystal and capacitor combination will differ from those of the crystal.

(b) Suppose the crystal is used in an oscillator where it operates in the "parallel resonant" mode, i.e., the frequency of oscillation is a frequency where the crystal impedance is inductive. How will the frequency of oscillation change if a capacitor is added in parallel with the crystal? Explain your reasoning and any underlying assumptions.

14. The crystal oscillator circuit shown in Figure 5.38 is called a "series mode" oscillator, because it will oscillate very close to the series resonant frequency of the crystal. Here the crystal grounds the base of the transistor at its series resonant frequency.

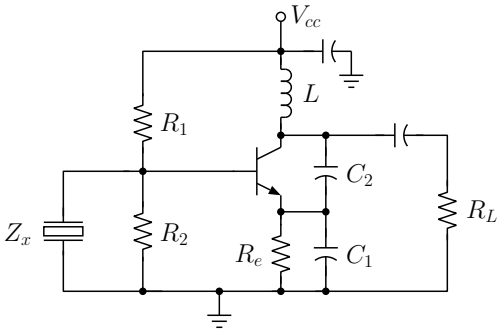


Figure 5.38: Series mode xtal oscillator derived from common-base Colpitts configuration.

The circuit will also oscillate if the crystal is replaced by an "AC" short circuit (e.g., a bypass capacitor). Use the negative resistance approach to study this oscillator. You may assume that the transistor can be modeled using the equivalent circuit in

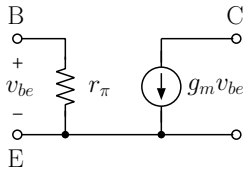


Figure 5.39: Simplified hybrid-pi model.

Figure 5.39. You can also assume that R_1 and R_2 can be neglected and that R_e is much larger than the reactance of C_2, so that it can be neglected as well. The coupling capacitors can be taken to be AC short circuits.

(a) First, remove the crystal and "look in" to the rest of the circuit. Solve for the input impedance and show that it is given by

$$Z_{in} = r_\pi + \frac{Z_1(Z_L + Z_2) + g_m r_\pi Z_1 Z_2}{Z_1 + Z_2 + Z_L} \tag{5.45}$$

where

$$\begin{aligned} Z_1 &= \frac{1}{j\omega C_1} \\ Z_2 &= \frac{1}{j\omega C_2} \\ Z_L &= \frac{j\omega L R_L}{j\omega L + R_L} \end{aligned} \tag{5.46}$$

(b) Solve for the frequency at which the circuit will oscillate (ω_o), if an AC short is connected from the base of the transistor to ground (instead of the crystal). You can assume that $R_L \gg \omega L$ at the frequency of oscillation.

(c) What condition must be satisfied in order to guarantee that oscillations will start?

15. Consider the oscillator circuit in Figure 5.40. The active device is modeled as a

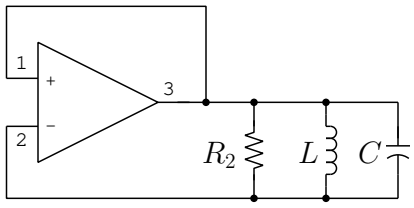


Figure 5.40: Oscillator circuit.

transconductance amplifier with input resistance R_1 as shown in Figure 5.41. (Note

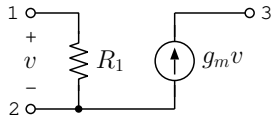


Figure 5.41: Transconductance amplifier with input resistance R_1

the important difference between this equivalent circuit and the simplified hybrid-pi model!)

(a) Find an expression for the potential frequency of oscillation, ω_o.

(b) Find the threshold value, ${g_m}^{min}$, which the transconductance must exceed in order for oscillation to occur.

16. Consider the oscillator circuit shown in Figure 5.42. In this problem, $V_{cc} = 3\,\text{V}$, $R_b = 33\,\text{k}\Omega$, $R_e = 100\,\text{k}\Omega$, $C_1 = 10\,\text{pF}$, $C_2 = 22\,\text{pF}$. Unlabeled capacitors are either coupling or bypass elements and have negligibly small impedance at the frequencies of interest. You may assume that the load resistor is disconnected from the circuit for parts (a)-(d).

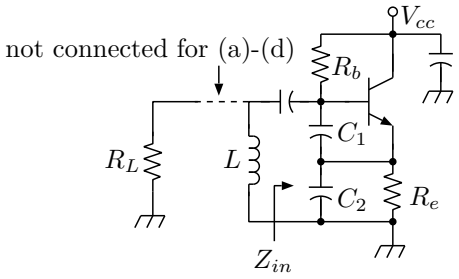


Figure 5.42: Load resistor is not connected for parts (a)-(d).

(a) Find an approximate value for the quiescent collector current, I_{CQ}. You may assume that $\beta = 100$.

(b) Find the transconductance, g_m, for the transistor in this circuit.

(c) Assume that the reactances of C_1 and C_2 are small enough such that r_π, R_e, and R_b can be ignored for small-signal analysis. In this case, if the inductor is removed the impedance Z_{in} looking in to the circuit at the point shown in the Figure will be

$$Z_{in} = -\frac{g_m}{\omega^2 C_1 C_2} + \frac{1}{j\omega \frac{C_1 C_2}{C_1 + C_2}}.$$

Ignore the load resistance, R_L, (i.e. assume that the load is disconnected) and calculate the inductance, L, required to set the potential frequency of oscillation in this circuit to 50 MHz.

(d) Suppose a lossy inductor is available with inductance equal to the value that you calculated in part (c). Calculate the minimum inductor Q (i.e., Q_L) required for oscillations to be sustained in this circuit.

(e) Now suppose that the actual inductor Q_L is 50. A load is placed on the oscillator by connecting R_L to the inductor through a coupling capacitor as shown by the dotted line in the figure. Calculate the smallest value of load resistance for which the circuit will still oscillate.

17. Design a tunable oscillator that covers the frequency range $10 - 20$ MHz using the Colpitts configuration shown in Figure 5.43 . The coil inductance $L = 2\mu H$ and

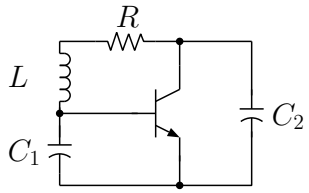


Figure 5.43: Colpitts oscillator with bias circuitry omitted.

the resistance $R = 4\Omega$. For simplicity, assume that the immittances associated with the transistor and the bias circuitry can be ignored. Furthermore, assume that the oscillator will be tuned with a single variable capacitor. For this problem, assume that C_2 is the variable capacitor and is adjustable within the limits $[C_{2,min}, C_{2,max}]$. As discussed in the supplementary notes, it is good practice to ensure that $C_1 \gg C_2$ to allow large collector voltage swing with relatively small base-emitter voltage swing. This allows the transistor to operate without being driven deep into saturation. For your design, use $C_1 = 10C_{2,max}$ in order to satisfy this constraint. Design your circuit so that the minimum value of the loop gain over the range 10-20 MHz is 2. For full credit you must specify the value of C_1, the minimum and maximum values for the variable capacitor (C_{2min} and C_{2max}), the transconductance, g_m, of the transistor, and the quiescent collector current I_{CQ}.

18. A crystal has packaging capacitance $C_o = 5$ pF and series and parallel resonant frequencies $f_s = 10\,\text{MHz}$ and $f_p = 10.002\,\text{MHz}$.

(a) Determine the inductance, L_1, and capacitance, C_1, of the elements in the motional arm of the equivalent circuit for the crystal.

(b) The crystal is used in a parallel-resonance oscillator in which the crystal resonates with an external load capacitance of 30 pF at the frequency of oscillation. The frequency of oscillation can be written as $f_o = 10.000\text{ MHz} + \delta f$. Specify δf to an accuracy of +/- 5 Hz.

19. Consider the crystal oscillator shown in Figure 5.44. Without the network consisting of the components labeled L and C, this circuit would oscillate near the fundamental (lowest) resonant frequency of the crystal, denoted by f_o, where $f_s < f_o < f_p$, and f_s,

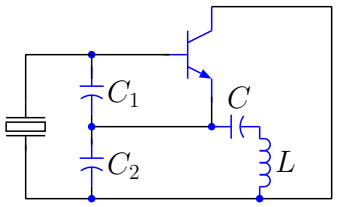


Figure 5.44: Crystal oscillator circuit for operation on 3rd overtone.

f_p are the lowest series and parallel resonant frequencies of the crystal. The purpose of the series LC network is to make it impossible for the circuit to oscillate at the fundamental frequency, and to allow the circuit to oscillate at the 3rd overtone, $3f_o$. If $f_o = 50$ MHz, $C_1 = 10$ pF, $C_2 = 30$ pF, $C = 500$ pF, find the range of values for L that will prevent the circuit from oscillating at f_o, and allow the circuit to oscillate at $3f_o$. For this problem, consider only the fact that the combination of C_2, C, and L must be capacitive in order for the circuit to oscillate at a frequency where the crystal is inductive. You do not need to ensure that the magnitude of the loop gain will be sufficient to actually cause oscillations to start and grow.

20. The circuit shown in Figure 5.45 is often used as an overtone XTAL oscillator. In this

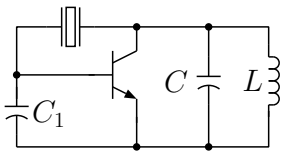


Figure 5.45:

problem, we consider how to design the circuit so that it oscillates at the 5'th overtone as a parallel-mode oscillator. Note that bias resistors and coupling/bypass capacitors are not shown and may be ignored. You may also ignore the internal immittances of the transistor.

(a) Consider only the XTAL in this part. Denote the series and parallel resonant frequencies associated with the 5'th overtone by $f_{s,5}$ and $f_{p,5}$, respectively. It is known that $f_{s,5} = 49.990$ MHz and $f_{p,5} = 50.030$ MHz and the packaging capacitance, C_o, is 5 pF. Find the external load capacitance, C_L, required to form a parallel resonant circuit with the XTAL at exactly 50.000 MHz.

(b) Write down a constraint that must be satisfied by the resonant frequency of the parallel LC branch in the circuit to allow the circuit to oscillate at the fifth overtone, and not at the fundamental or third overtone. Denote the desired frequency of oscillation by f_o. The XTAL's "fundamental" resonances are near $f_o/5$ and the "third overtone" resonances near $3f_o/5$.

(c) Write an equation in terms of C_L, C_1, L, and C which must be satisfied (along with the constraint derived in part b.) to set the lowest potential frequency of oscillation to exactly 50.000 MHz ($f_o = 50.000$ MHz). You do not need to solve the equation.

21. Consider the Colpitts oscillator shown in Figure 5.46.

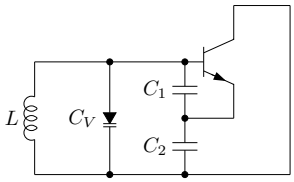


Figure 5.46: Voltage-controlled oscillator (VCO).

(a) Find an expression for the frequency of oscillation, ω_o. You may assume that the loop gain of the circuit is large enough to cause oscillation to start. You may neglect the transistor immittances, so that your result will be in terms of only the four parameters L, C_1, C_2, C_V.

(b) Denote the minimum and maximum tuning-diode capacitances by $C_{V,min}$ and $C_{V,max}$ and the capacitor tuning ratio by $r_C = \frac{C_{V,max}}{C_{V,min}}$. Denote the minimum and maximum frequencies of oscillation by $\omega_{o,max}$ and $\omega_{o,min}$ and the oscillator tuning ratio by $r_o = \frac{\omega_{o,max}}{\omega_{o,min}}$. Finally, denote the series combination of C_1 and C_2 by C', i.e. $C' = \frac{C_1 C_2}{C_1 + C_2}$. Find an expression for r_o in terms of r_C, C', and $C_{V,max}$ only.

(c) Find a numerical value for the tuning ratio r_o if $C' = 2C_{V,max}$ and $r_C = 4$.

Chapter 6

Impedance Matching Networks

6.1 Impedance Matching for Maximum Power Transfer

In this section we review the motivation for impedance matching and introduce important concepts which will be used in later chapters. Let us first illustrate the basic principles of impedance matching for maximum power transfer. In Figure 6.1 a source with impedance $Z_S = R_S + j\,X_S$ is connected to a load $Z_L = R_L + j\,X_L$. The peak voltage of the source (assumed to be sinusoidal) is V_S:

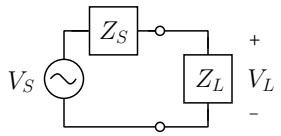


Figure 6.1: A voltage source driving an arbitrary load

The time-averaged real power delivered to the load can be written as

$$P_L = \frac{1}{2}\mathrm{Re}\{V_L I_L^*\} \tag{6.1}$$

where V_L and I_L are the voltage across and current through the load impedance, respectively. Note that the factor $\frac{1}{2}$ is present because V_L and I_L are "peak" phasors,i.e., phasors whose magnitude is the peak value of the sinusoidal time function. The current through the load is

$$I_L = \frac{V_L}{Z_L} \tag{6.2}$$

so the time-averaged real power is

$$\begin{aligned} P_L &= \frac{1}{2}|V_L|^2 Re\{\frac{1}{{Z_L}^*}\} \qquad (6.3)\\ &= \frac{1}{2}|V_L|^2 \frac{R_L}{|Z_L|^2} \end{aligned}$$

This can be written in terms of the source voltage V_S as follows:

$$\begin{aligned} P_L &= \frac{1}{2}|V_S|^2 \frac{R_L}{|Z_L + Z_S|^2} \\ &= \frac{1}{2}|V_S|^2 \frac{R_L}{(R_L + R_S)^2 + (X_L + X_S)^2} \end{aligned} \tag{6.4}$$

By studying Equation 6.4 we can determine what conditions are required to maximize the power delivered to the load. First we will assume that the source voltage and source impedance are set at the outset, i.e., they are not under our control. This will usually be the case. The problem is then to choose R_L and X_L to maximize P_L. We will restrict the solution to values of R_L that are greater than or equal to 0 because values of R_L that are < 0 would require that the load contain an active device, whereas we are considering only passive load terminations. The solution to this simple exercise is

$$Z_L = Z_S^* \tag{6.5}$$

This result states that the time-average real power delivered to a load is maximized when the load impedance is equal to the complex conjugate of the source impedance. Using Equation 6.5 in Equation 6.4, we obtain an expression for the maximum power that the source can deliver to an external passive load. This power is referred to as the power available from the source, P_{avs}:

$$\begin{aligned} P_{avs} &\equiv P_L|_{Z_L = Z_S^*} \\ &= \frac{|V_S|^2}{8R_S} \end{aligned} \tag{6.6}$$

The concept of *available power* is often used in the radio frequency literature. For example, the output power level of RF signal generators is usually specified by stating the available power in *dBm*, i.e., "decibels referred to 1 mW." A power level of 6 dBm indicates a power 6 dB higher than 1 mW, or approximately 4 mW. It is important to realize that when a signal generator is configured for a given output power level, that power level is the power available from the generator, P_{avs}, which is the power that the generator will deliver to a conjugately matched load. Signal generators commonly have source impedances of $50\,\Omega$ (sometimes $75\,\Omega$), i.e., the generator will deliver the rated power only to a $50\,\Omega$ load. With a different load impedance the power delivered to the load will be less than the rated value.

6.1.1 Mismatch Factor

The degree to which the actual power delivered to an arbitrary load is smaller than the available power can be quantified in terms of a *mismatch factor*, MF, a quantity that depends on the degree of impedance mismatch between the source and load. For an arbitrary load impedance Z_L the mismatch factor is defined as the ratio of actual delivered power to available power:

$$\begin{aligned} MF &= \frac{P_L}{P_{avs}} \\ &= \frac{4R_S R_L}{(R_S+R_L)^2+(X_S+X_L)^2} \\ &= \frac{4R_S R_L}{|Z_S+Z_L|^2} \end{aligned} \tag{6.7}$$

Mismatch factor is a real number and $0 \leq MF \leq 1$. For given source and load impedances, the mismatch factor tells us what fraction of the available power will be delivered to the load. Since the mismatch factor is a ratio of two power levels, it makes sense to express the quantity in decibels. The *mismatch loss*, ML, in decibels is defined as

$$ML = -10 \log MF. \tag{6.8}$$

6.1.1.1 Example - Mismatch Factor and Mismatch Loss

Suppose a $50\,\Omega$ signal generator has available power of 1 mW. The generator is to drive a load impedance of $250 + j100\,\Omega$. What is the power delivered to the load?

The problem could be solved by computing the power delivered to the load using Equation 6.1. Another approach is to compute the mismatch factor from Equation 6.7. This gives a mismatch factor $MF = 0.5$, so $P_L = P_{avs}MF = (1\,\text{mW})0.5 = 0.5\,\text{mW}$. Alternatively, we can express the powers in dBm and use the mismatch loss in dB. The mismatch loss $ML = -10\log(0.5) = 3\,\text{dB}$. The power available from the source is 1 mW, or $10\log\frac{1\,\text{mW}}{1\,\text{mW}} = 0\,\text{dBm}$. The power delivered to the load is $P_L = P_{avs} - ML = 0\,\text{dBm} - 3\,\text{dB} = -3\,\text{dBm}$. Note that a power level of $-3\,\text{dBm}$ is equal to 0.5 mW.

6.1.2 Properties of Lossless Impedance Matching Networks

We have illustrated how impedance matching (or mis-matching) influences the transfer of power from a source to a load. In many applications it is desirable to maximize the transfer of power from the source to the load. This can be achieved by using a lossless 2-port network inserted between the source and the load. The purpose of the *matching network* is to transform the load impedance, Z_L, into Z_S^* at the input terminals, thereby permitting the source to deliver all of its available power to the network. If a matching network is lossless, then all of the power that is delivered to the network must be delivered to the load. This simple concept has implications that may not be obvious at first glance.

Consider a source and load connected through a lossless matching network as shown in Figure 6.2. As already noted, the matching network transforms the load impedance Z_L into Z_S^* at the input terminals of the network. Since the source looks into the conjugate of its impedance, it delivers all of its available power P_{avs} to the network. Because no power is dissipated in a lossless network, all of this power is delivered to the load. Now the power available at the output of the lossless matching network must be the same as the power available from the source, and we have already stated that all of this available power is delivered to the load. Therefore the load must be conjugately matched to the output of the matching network, which means that the impedance at the output of the network (as seen by the load) is Z_L^*. Thus a lossless matching network has the property of providing a simultaneous conjugate match at the input and output ports.

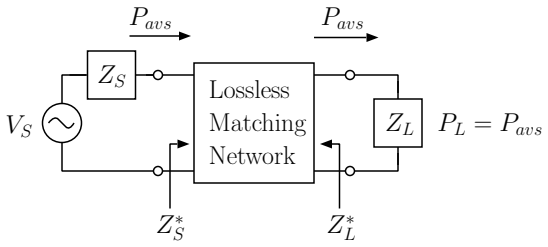


Figure 6.2: Source and load coupled by a lossless matching network

Another way to look at this is to note that the original source plus the matching network can be viewed as a new source with available power P_{avs} and source impedance Z_L^*. Adding a lossless network to the output of a source does not change the available power from the source; it only changes the source impedance. Thus, we can think of the matching network as (i) a network that transforms the load impedance into Z_S^* or (ii) a network that transforms the source impedance into Z_L^* without changing the available power of the original source. The network can be designed either way.

Most lumped-element matching networks are versions of ladder networks as shown in Figure 6.3. The ladder-type matching network in Figure 6.3 is assumed to be lossless; the

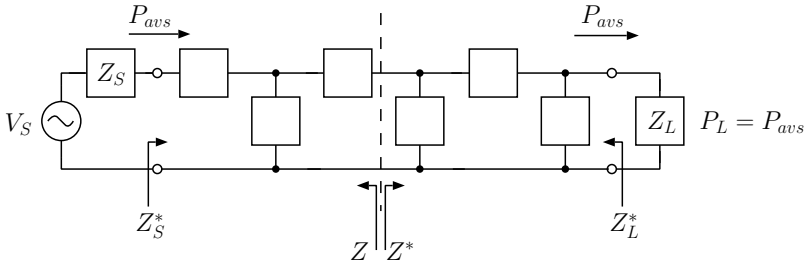


Figure 6.3: Ladder-type matching network

blank boxes represent purely reactive circuit elements. Now suppose that the network is "broken" at the dashed line. The part of the network to the left of the line can be thought of as a source. It will have the same available power as that of the original source, P_{avs}, and some source impedance, Z. The part of the network to the right of the dashed line is a load for this source and must receive all of the available power (since that power is ultimately transferred to Z_L). Thus, a conjugate match must exist at the junction defined by the dashed line. Clearly, the dashed line could have been drawn anywhere within the lossless network and the same argument would hold. **The conclusion is: A circuit consisting of a source connected to a load through a lossless matching network can be broken at any point between the source and the load and a conjugate match must exist between the two sides of the circuit**.

6.2 Impedance Matching with Lossless L-networks

6.2.1 Resistive Terminations

Figure 6.4 shows two resistances to be matched with a lossless L-network. The goal is to transform R_1 to R_2 at one frequency. The unknown reactances X_s and X_p are easily found with the use of a parallel-to-series transformation as in Figure 6.5.

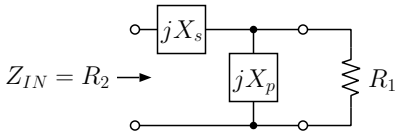


Figure 6.4: L-network which transforms R_1 into R_2

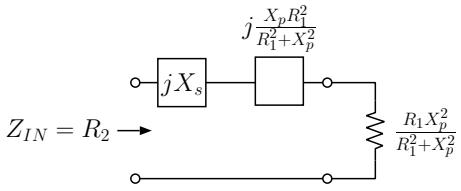


Figure 6.5: Parallel-to-series transformation

To make the input impedance equal to R_2, we choose

$$X_s = -\frac{X_pR_1^2}{R_1^2 + X_p^2} \tag{6.9}$$

$$R_2 = \frac{R_1X_p^2}{R_1^2 + X_p^2}$$

Solving for X_p and X_s gives

$$X_p = \pm R_1\sqrt{\frac{R_2}{R_1 - R_2}} \tag{6.10}$$

$$X_s = \mp\sqrt{R_2R_1 - R_2^2} \tag{6.11}$$

The solutions yield real values for X_p and X_s (and hence purely reactive L-network components) only if $R_1 > R_2$. This leads to a rule for using a lossless L-network to match two resistances:

The shunt arm of the L-network is connected across the larger of the two resistances.

Also note that there are two possible solutions for the resistive matching problem corresponding to the upper and lower signs in Equation 6.10 and Equation 6.11. These solutions are referred to as "low-pass" and "high-pass" solutions. The circuit configurations for the two solutions are in shown Figures 6.6 and 6.7.

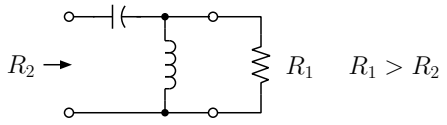


Figure 6.6: High-pass L-network

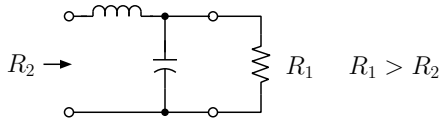


Figure 6.7: Low-pass L-network

The high-pass and low-pass designations refer to the behavior of the network frequency response function at frequencies away from the design frequency.

6.2.2 "Q" of an L-network

Referring to Figure 6.5, it is apparent that the transformed L-network looks like a series-resonant circuit. At the design frequency the two reactances have equal magnitudes and opposite signs, so that the net series reactance is zero. Treating this circuit in the same manner as a resonant RLC network, the Q of the network is the magnitude of one of the series reactances divided by the series resistance:

$$Q = \left| \frac{X_p R_1^2/(R_1^2 + X_p^2)}{R_1 X_p^2/(R_1^2 + X_p^{2)}} \right| \tag{6.12}$$

$$\tag{6.13}$$

$$= \left| \frac{R_1}{X_p} \right| \tag{6.14}$$

$$\tag{6.15}$$

$$= \sqrt{\frac{R_1}{R_2} - 1} \tag{6.16}$$

This result shows that the Q of the L-network is determined by the terminating resistances R_1 and R_2, and, therefore, cannot be chosen independently. This suggests that the bandwidth of the matching network depends only on the ratio R_1/R_2.

A word of caution is in order. We have defined the Q of the L-network by analogy with the series RLC network. This analogy works well only when $R_1 \gg |X_p|$ or, equivalently, when $Q \gg 1$. In such cases the Q can be used to predict the 3 dB bandwidth of the network's voltage transfer function. For moderate or small values of Q the expression

$$Q = \frac{R_1}{|X_p|} = \sqrt{\frac{R_1}{R_2} - 1}$$

is still valid, but a simple relationship between Q and bandwidth does not exist, since the series reactance that results from the parallel-to-series transformation has a different frequency dependence from that of a simple inductor or capacitor. Thus an L-network does not behave exactly like a series RLC circuit. Only when the Q is very large will the equivalent series reactance behave approximately like a capacitor or inductor.

6.2.3 Summary: L-network design equations

The design of an L-network can be summarized as follows. If the terminating resistances are denoted by R_{big} and R_{small} (where $R_{big} > R_{small}$), then the design equations can be written in terms of the network Q,

$$Q = \sqrt{\frac{R_{big}}{R_{small}} - 1}, \tag{6.17}$$

as

$$X_p = \pm R_{big}/Q \qquad X_s = \mp R_{small} Q. \tag{6.18}$$

When the upper signs are chosen, the resulting network is of highpass type. The lower signs give a lowpass network. The L-network will be oriented such that the parallel arm is in shunt with R_{big}.

6.2.3.1 Example - Matching resistive source and load with a low-pass L-network

Match a 100 Ω source to a 25 Ω load with a lossless L-network having a low-pass topology. Since the series arm of the L-network connects to the smaller of the two resistances, the L-

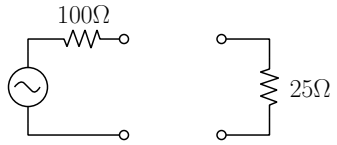


Figure 6.8: Example for L-net matching

network will be oriented as shown in Figure 6.9. A capacitor and inductor have been chosen for the shunt and series elements, respectively, because the problem statement specified a low-pass network.

$$\begin{aligned} Q &= \sqrt{\tfrac{R_{big}}{R_{small}} - 1} = \sqrt{\tfrac{100}{25} - 1} = \sqrt{3} \\ X_p &= X_C = -\tfrac{R_{big}}{Q} = -\tfrac{100}{\sqrt{3}} = -57.7\,\Omega \\ X_s &= X_L = QR_{small} = \sqrt{3}\,25 = 43.3\,\Omega \end{aligned}$$

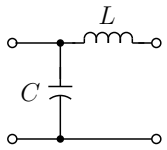


Figure 6.9: L-network orientation

The completed network is shown in Figure 6.10. It is interesting to compare the power

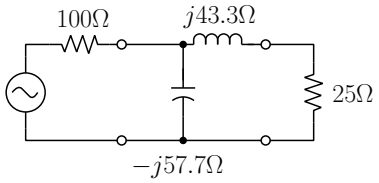


Figure 6.10: Completed lossless L-network with low-pass topology

delivered to the load with and without the matching network in place. Suppose the peak voltage of the source is 1 Volt. Then the power delivered to the load without the matching network would be

$$P_{out} = \frac{1}{2}\frac{|V_{out}|^2}{25} = \frac{1}{50}\left(\frac{1}{5}\right)^2 = 0.8\,\text{mW} \tag{6.19}$$

With the matching network the source "sees" 100 Ω; therefore, the power delivered to the matching network is

$$P_{in} = \frac{1}{2}\frac{(1/2)^2}{100} = \frac{1}{800} = 1.25\,\text{mW} \tag{6.20}$$

Since the matching network is lossless, all of this power will be delivered to the load. Thus, with the matching network, $P_{out} = 1.25$ mW. The improvement gained is

$$10\log(1.25/.8) \simeq \underline{2\,\text{dB}} \tag{6.21}$$

6.2.4 Matching Complex Loads with a Lossless L-network

So far, only purely resistive source and load terminations have been considered. When complex source and loads are involved, there are two basic conceptual approaches that can be used:

1. Absorption - "absorb" the source or load reactance into the matching network.

2. Resonance - series or parallel resonate the source or load reactance at the frequency of interest.

These approaches will be illustrated by example in the following sections.

6.2.4.1 Example - Absorption

Absorption will be illustrated with an example. Suppose that it is necessary to match the source and load shown in Figure 6.11 at 100 MHz with a lossless L-network.

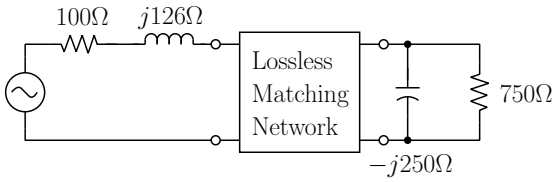


Figure 6.11: L-network with complex source and load

Absorption is applied by lumping the source and load reactances into the series and parallel reactances of the matching network, as shown in Figure 6.12. The lumped reactances

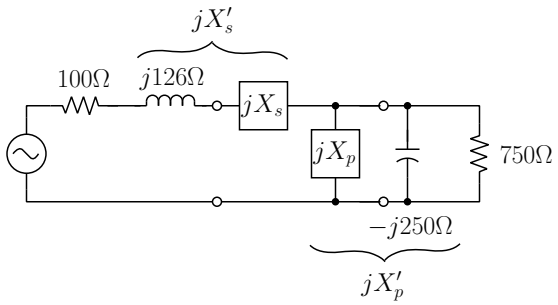


Figure 6.12: Absorbing the complex part of the load impedance into the network

X'_s and X'_p can be found by using the design equations for matching between two resistive terminations:

$$\begin{aligned} Q &= \sqrt{750/100 - 1} &&= 2.55 \\ X'_s &= \pm 100\,(2.55) &&= \pm 255 \\ X'_p &= \mp 750/2.55 &&= \mp 294.1 \end{aligned} \tag{6.22}$$

The lumped reactances can be written in terms of the reactances associated with the source and load, and the L-network reactances:

$$X'_s = X_s + 126, \tag{6.23}$$

$$X'_p = \frac{-250\,(X_p)}{-250 + X_p}. \tag{6.24}$$

Thus, values of X_s and X_p can be obtained:

$$\begin{aligned} X_s &= X_s' - 126 = \pm 255 - 126 = \begin{cases} 129\Omega \\ -381\Omega \end{cases} \\ X_p &= \frac{250X_p'}{250+X_p'} = \begin{cases} 1667.2\Omega \\ 135.1\Omega \end{cases} \end{aligned} \tag{6.25}$$

The two solutions found so far are shown in Figure 6.13.

Figure 6.13: Two L-network solutions.

The element values (in μH and pF) were obtained from the calculated reactances using the design frequency of 100 MHz. It should be noted that the two solutions found so far are not the only possibilities for this particular source and load. This becomes apparent if series-to-parallel transformations are applied to the source and load. After transformation, the source and load look like Figure 6.14.

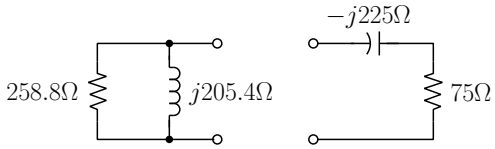


Figure 6.14: Source and load after application of series-to-parallel transformation

For simplicity, the Norton equivalent current source has been omitted from the source representation. Note that the smaller of the two resistances is now on the load side and hence it is possible to match the source and load using L-networks oriented as in Figure 6.15.

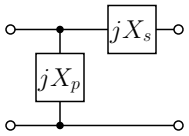


Figure 6.15: L-network reoriented for smaller resistance on load side

The values of X_s and X_p could be found using absorption. Alternatively, the resonance concept can be used. For illustration, the resonance concept will be used to find the remaining L-network solutions.

6.2.4.2 Example - Resonance

Continuing with the example from the previous section, the resonance concept will be employed to find two more L-network solutions for the source and load shown in Figure 6.11. After transforming the source from series to parallel form, and the load from parallel to series form, the source and load are represented as in Figure 6.14. To apply the resonance concept, the source and load are augmented with reactances that resonate with the source and load reactances as shown in Figure 6.16.

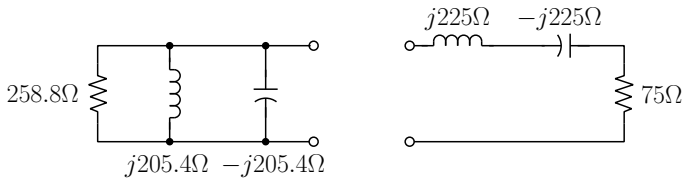


Figure 6.16: Transformed source and load augmented with resonating reactances

After resonating the source and load in this manner, the new source and load impedances are purely real (resistive) at the design frequency. An L-network is then designed to match these two resistances, i.e.,

$$\begin{aligned} Q &= \sqrt{\frac{258.76}{75} - 1} &&= 1.57 \\ X'_s &= \pm 75\,(1.57) &&= \pm 117.8 \\ X'_p &= \mp 258.76\,/1.57 &&= \mp 164.8 \end{aligned} \tag{6.26}$$

Now the circuit can be drawn as shown in Figure 6.17. To complete the design, the resonating

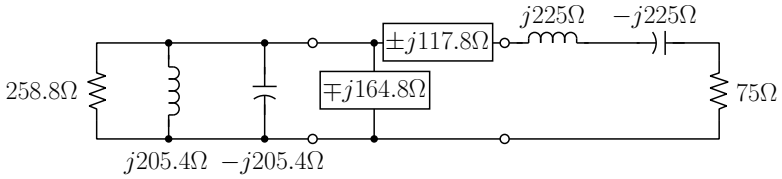


Figure 6.17: L-net with resonating reactances

reactances must be incorporated into the matching network. For example, when the upper signs are chosen, the net parallel reactance will be $-j205.4\Omega || -j164.8\Omega = -j91.4\Omega$. The net series reactance will be $j117.8\Omega + j225\Omega = j342.8\Omega$. The final solutions are shown in Figure 6.18.

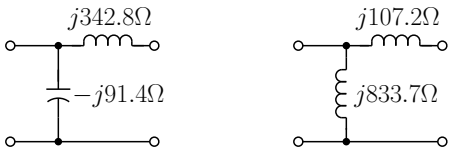


Figure 6.18: Resonating reactances incorporated into matching network

Thus, we have found four possible solutions that can be obtained using a lossless L-network.

The example considered here allowed 4 possible L-network solutions because transforming the source and load caused R_{big} and R_{small} to swap positions. This will not always be the case. Thus, for some source and load combinations, there will be only two L-network solutions, and in other cases there will be four solutions.

6.3 Harmonic Attenuation in Lossless Matching Networks Using Traps

In certain applications it is necessary to provide a match at some frequency ω_o and to attenuate one or more other frequencies. Typically the undesired frequencies are harmonics of ω_o, i.e., $2\omega_o$, $3\omega_o$, etc.; however, they could also be subharmonic or even unrelated to ω_o.

We will concentrate on the case where the undesired frequencies are harmonics. Generalization to other cases is straightforward. The basic idea is to first design a lossless matching network to provide a match at the desired frequency (ω_o). This matching network might consist of an L-, T-, or Pi-network. Next, one or more series elements of the network are replaced with parallel L-C-networks. Alternatively, or in addition, the shunt elements of the network are replaced with series L-C-networks. The replacement (series or parallel L-C) elements are designed to have the same reactance as the original elements at ω_o; however, they are designed to be resonant at the undesired harmonic frequency. As an example, consider Figure 6.19 which shows a matching network with a match at $f_o = 10^7/2\pi$ Hz.

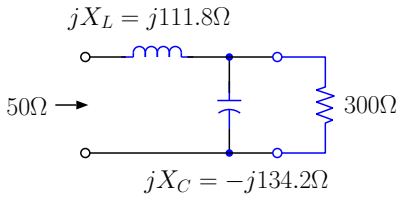


Figure 6.19: $L = 11.18\,\mu$H, $C = 745.2\,$pF

Suppose it is desired to "trap" the second harmonic $2f_o$ in the shunt arm. The shunt arm would be replaced with a series L-C that has reactance $-134.2\,\Omega$ at f_o and is series

resonant (looks like a short circuit) at $2f_o$. The design equations are

$$2\,\omega_o = 1/\sqrt{LC} \tag{6.27}$$

$$\omega_o\,L - \frac{1}{\omega_o C} = -134.2 \tag{6.28}$$

The solution is

$$L = 4.47\,\mu\text{H} \tag{6.29}$$

$$C = 558.9\,\text{pF}$$

The new network looks like Figure 6.20.

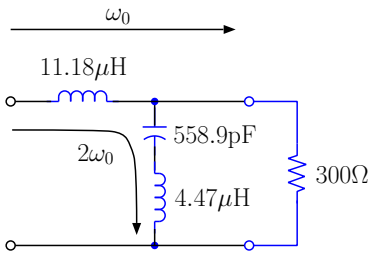


Figure 6.20: Network with "trapped" second harmonic

The series trap acts to shunt the component at frequency $2f_o$ to ground.

The third (or any other) harmonic could be trapped in the series arm. The inductor would be replaced with a parallel LC whose element values are found from

$$3\omega_o = \frac{1}{\sqrt{LC}} \tag{6.30}$$

$$-\frac{1}{\omega_o\,L} + \omega_o\,C = -\frac{1}{111.8} \tag{6.31}$$

The solution is

$$C = 111.8\,\text{pF} \tag{6.32}$$

$$L = 9.94\,\mu\text{H}$$

Figure 6.21 shows the network with both traps installed. The network will look like an open circuit to the $3f_o$ component.

It should be noted that this approach will work only when applied to *capacitive* shunt elements and *inductive* series elements if frequencies *higher* than ω_o are to be trapped. Thus, the original network must have a low-pass topology. The student should convince himself or herself that such is the case. On the other hand, if frequencies smaller than ω_o are to be trapped, then the approach can be employed only with inductive shunt elements and capacitive series elements.

The trapping approach is also useful for reducing the feed-through of nonharmonic frequency components. If the frequencies of the undesired components are very close to the desired frequency, it may not be possible to build an effective trap, because the Q of the series or parallel L-C circuits will be too high to be realized.

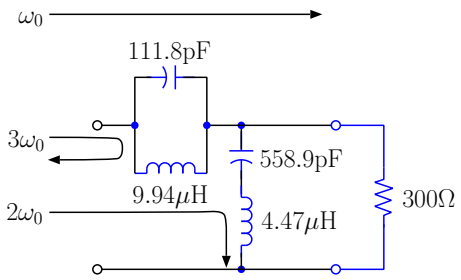


Figure 6.21: Network with second and third harmonics trapped

6.4 Three-element Matching Networks

The L-network does not give the designer freedom to choose the Q (bandwidth) or phase shift of the matching network. The addition of a third matching element makes it possible to design for a match and a specified phase shift or Q. The following section describes a procedure for designing 3 element matching networks with specified Q. Then, a general procedure allowing for specified attenuation and phase shift will be discussed.

6.4.1 Design of Pi- and T-networks for Specified Bandwidth (Q)

First consider the Pi-network with a low-pass topology. In addition, we assume that absorption has been used so that the shunt reactance of the source or load impedance is included in the Pi-network. The problem is then reduced to matching between two resistive terminations as shown in Figure 6.22 (where it is assumed that $R_2 > R_1$).

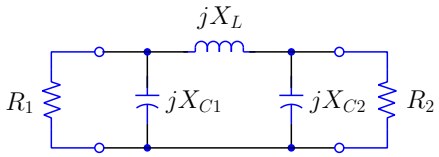


Figure 6.22: Matching two resistive terminations where $R_2 > R_1$

The Pi-network can be thought of as two back-to-back L-networks that act to match both R_1 and R_2 to a "virtual resistance" R_v as shown in Figure 6.23.

Because the series arms of both L-networks are connected to R_v, it is clear that R_v is smaller than R_1 and R_2. Define the Q's of the two L-networks to be Q_1 and Q_2 where

$$Q_1 = \sqrt{\frac{R_1}{R_v} - 1}\,, \quad \text{and} \quad Q_2 = \sqrt{\frac{R_2}{R_v} - 1} \tag{6.33}$$

We are assuming that $R_2 > R_1$, and therefore Q_2 will be larger than Q_1. For most practical purposes the Q of the Pi-network can be approximated by Q_2. This is especially true if $R_2 \gg R_1$. If R_2 is only slightly larger than R_1, then the overall Q of the network will be somewhat larger than Q_2. The design procedure follows:

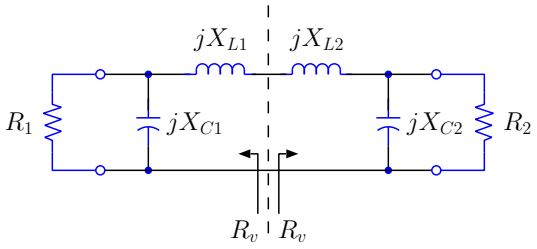


Figure 6.23: The Pi-network as 2 back-to-back L-networks

1. Determine the required Q of the matching network by considering the required bandwidth, BW, and center frequency, f_o ($Q = f_o/BW$). This Q is taken to be equal to Q_2, and thus the virtual resistance, R_v, is determined. **Note that R_v must be smaller than R_1, and therefore the Pi-network can only be used to obtain a larger Q than would have been provided by the simpler L-network**. Also note that the relationship $BW = f_o/Q$ is only exactly true for a simple parallel or series RLC circuit. Thus, the actual bandwidth of your circuit may be different from the specified design value. If a particular design requires that the bandwidth be precisely determined, it is a good idea to simulate the performance of the matching network using a computer-aided design program in order to verify that the performance will be satisfactory.

2. Once R_v is found, the values of X_{C1}, X_{L1}, X_{C2}, and X_{L2} can be calculated using the previously derived formulas for L-network matching.

The design procedure can be summarized by Equations 6.34, where Q_2 is determined by the desired bandwidth. You should verify that these equations result from steps (1) and (2):

$$X_{C2} = -\frac{R_2}{Q_2} \tag{6.34}$$

$$X_{C1} = -\sqrt{\frac{R_1 R_2}{(Q_2^2+1) - \frac{R_2}{R_1}}} \tag{6.35}$$

(6.36)

(6.37)

$$X_L = \frac{R_2 Q_2 + R_2\sqrt{\frac{R_1}{R_2}(Q_2^2+1) - 1}}{Q_2^2+1} \tag{6.38}$$

The Pi-network is most useful for matching when the values of R_1 and R_2 are not too small. If R_1 and R_2 are small, the virtual resistance will be even smaller, and the capacitor values will turn out to be impractically large. If either terminating resistance is significantly less than 50 Ω, the T-network will usually be a more practical choice. One possible T-network is the band-pass case shown in Figure 6.24.

As before, we can think of this network as two back-to-back L-networks as shown in Figure 6.25.

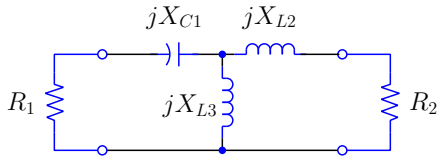


Figure 6.24: Band-pass T-network

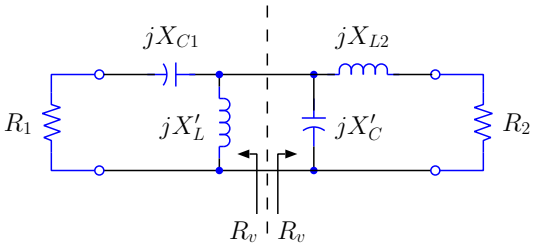


Figure 6.25: Band-pass T-network as 2 back-to-back L-networks

The Q's of the two L-networks are

$$Q_1 = \sqrt{\frac{R_v}{R_1} - 1} \quad \text{and} \quad Q_2 = \sqrt{\frac{R_v}{R_2} - 1} \tag{6.39}$$

Note that since $R_v > R_1$ and $R_v > R_2$, this network will have a larger Q than a single L-network that matches R_1 to R_2. The overall Q of the network is set by Q_1 since we assume $R_2 > R_1$. The design formulas are

$$X_{C1} = -R_1 Q_1 \tag{6.40}$$

$$X_{L2} = R_2 \sqrt{\frac{R_1}{R_2}(Q_1^2 + 1) - 1} \tag{6.41}$$

$$\tag{6.42}$$

$$X_{L3} = \frac{R_1(Q_1^2 + 1)}{Q_1 - \sqrt{\frac{R_1}{R_2}(Q_1^2 + 1) - 1}} \tag{6.43}$$

Note carefully that the two elements X'_L and X'_C that appear in the back-to-back L-networks are combined into a single element, X_{L3}, in the T-network. In practice this is not necessary and, in fact, may not be desirable. Whether or not the elements are combined makes no difference at the design frequency, but it may make a significant difference at frequencies well removed from the design frequency. The different possibilities are best examined using a CAD program.

You should be able to derive similar formulas for the other possible topologies, e.g., band-pass Pi, low-pass T, etc.

6.4.2 Matching Two Resistive Terminations with Specified Attenuation and Phase Shift

In this section we will consider a more general type of matching network than was considered in section 6.2. Specifically, we allow the matching element impedances to be complex in the initial development of our solution. This will allow for a solutions with specified attenuation and phase shift. Then we will specialize to the cases where the elements are purely reactive (lossless networks) and purely resistive (lossy networks or attenuators). Note that the terminating impedances are assumed to be real (resistive).

6.4.2.1 Pi-network with specified attenuation and phase shift

Referring to Figure 6.26 we make the following assumptions:

Y_1, Y_2 are assumed to be real.

Y_A, Y_B, Y_C may be complex.

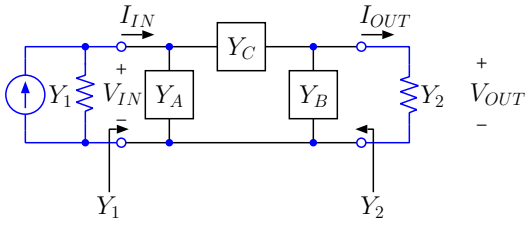


Figure 6.26: Matching two resistive terminations with a Pi-network

The conditions required for a match can be summarized as follows:

1. Must see Y_1 looking in from the left when terminated with Y_2 on the right. Thus,

$$Y_1 = Y_A + \frac{(Y_2 + Y_B)Y_C}{Y_2 + Y_B + Y_C} \tag{6.44}$$

2. Must see Y_2 looking in from the right when terminated with Y_1 on the left. Thus,

$$Y_2 = Y_B + \frac{(Y_1 + Y_A)Y_C}{Y_1 + Y_A + Y_C} \tag{6.45}$$

So far, we have 2 equations but 3 unknowns. The third equation can be used to specify either the Q of the network (and therefore, the bandwidth) or the phase shift and attenuation. For now, we consider the latter. The third equation can be written in the form:

$$e^{\alpha+j\beta} = \sqrt{\frac{V_{IN}\, I_{IN}}{V_{OUT}\, I_{OUT}}} \tag{6.46}$$

Equation 6.46 can be interpreted as follows. Since Y_1 and Y_2 are assumed to be real, the input voltage and current will be in phase and so will the output voltage and current.

Denoting the phase of the input voltage (current) by θ_{IN} and the phase of the output voltage (current) by θ_{OUT}

$$e^{\alpha+j\beta} = \sqrt{\frac{|V_{IN}|\,|I_{IN}|}{|V_{OUT}|\,|I_{OUT}|}}\, e^{j(\theta_{IN}-\theta_{OUT})} \tag{6.47}$$

$$= \sqrt{\frac{P_{IN}}{P_{OUT}}}\, e^{j(\theta_{IN}-\theta_{OUT})}$$

Thus

$$\beta = \theta_{IN} - \theta_{OUT} \tag{6.48}$$

and

$$\alpha = \frac{1}{2}\ln\frac{P_{IN}}{P_{OUT}} \tag{6.49}$$

Therefore, β is the phase shift of the network and α is the power attenuation expressed in nepers.

Equations 6.44, 6.45 and 6.46 must be solved for the unknowns Y_A, Y_B and Y_C. Note that Equation 6.46 can be written in terms of the unknowns as shown below. Define $\theta = \alpha + j\beta$, then

$$e^{\theta} = \sqrt{\frac{V_{IN}\,I_{IN}}{V_{OUT}\,I_{OUT}}} = \frac{V_{IN}}{V_{OUT}}\sqrt{\frac{I_{IN}/V_{IN}}{I_{OUT/V_{OUT}}}} \tag{6.50}$$

$$\tag{6.51}$$

$$= \frac{V_{IN}}{V_{OUT}}\sqrt{\frac{Y_1}{Y_2}} \tag{6.52}$$

$$\tag{6.53}$$

$$e^{\theta} = (1 + Y_C^{-1}(Y_2 + Y_B))\sqrt{\frac{Y_1}{Y_2}} \tag{6.54}$$

Equations 6.44, 6.45 and 6.46 can now be solved. The result is

$$Y_C = \frac{\sqrt{Y_1 Y_2}}{\sinh\theta} \tag{6.55}$$

$$Y_B = \frac{Y_2}{\tanh\theta} - Y_C \tag{6.56}$$

$$Y_A = \frac{Y_1}{\tanh\theta} - Y_C \tag{6.57}$$

6.4.2.2 T-network with specified attenuation and phase shift

Similar considerations apply to using a T-network to match two resistive terminations as shown in Figure 6.27.

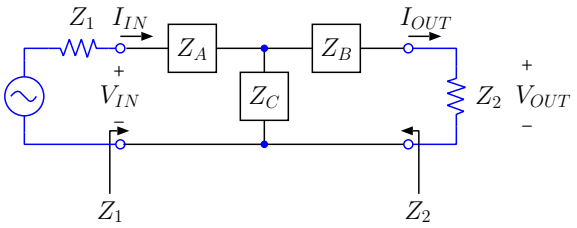


Figure 6.27: Matching two resistive terminations with a T-network

The equations that ensure a simultaneous conjugate match at both ports are

$$Z_1 = Z_A + \frac{Z_C(Z_B + Z_2)}{Z_C + Z_B + Z_2} \tag{6.58}$$

$$Z_2 = Z_B + \frac{Z_C(Z_A + Z_1)}{Z_C + Z_A + Z_1} \tag{6.59}$$

The third equation is the same as before,

$$e^{\alpha+j\beta} = e^{\theta} = \sqrt{\frac{V_{IN}\, I_{IN}}{V_{OUT}\, I_{OUT}}} \tag{6.60}$$

which can be written

$$e^{\theta} = \frac{I_{IN}}{I_{OUT}}\sqrt{\frac{Z_1}{Z_2}} \tag{6.61}$$

The solutions for Z_A, Z_B and Z_C are

$$Z_C = \frac{\sqrt{Z_1\, Z_2}}{\sinh\theta} \tag{6.62}$$

$$Z_B = \frac{Z_2}{\tanh\,\theta} - Z_C \tag{6.63}$$

$$Z_A = \frac{Z_1}{\tanh\,\theta} - Z_C \tag{6.64}$$

The interpretation of $\theta = \alpha + j\beta$ is the same for the T-network as it was for the Pi-network, that is,

$$\begin{aligned} \beta &= \theta_{IN} - \theta_{OUT} \quad \Rightarrow \quad \text{voltage (current) phase shift in radians} \\ \alpha &= \tfrac{1}{2}\ln\tfrac{P_{IN}}{P_{OUT}} \quad \Rightarrow \quad \text{power attenuation in nepers} \end{aligned} \tag{6.65}$$

Remember that these discussions have assumed that Y_1 and Y_2 (or Z_1 and Z_2) are real (resistive). Complex loads are handled by incorporating the reactive part of the termination into the matching network using either resonance or absorption as discussed in section 6.2.

6.4.3 Design of Lossless Pi- and T- Matching Networks with Specified Phase Shift

The solutions found so far allow the designer to specify both phase shift and attenuation of the network. In practice one is usually concerned with one of the special cases: (i) Y_A, Y_B, Y_C or Z_A, Z_B, Z_C are purely reactive — this corresponds to the lossless matching network; (ii) Y_A, Y_B, Y_C orZ_A, Z_B, Z_C are purely resistive — this corresponds to a lossy network. Networks of type (ii) are often used to provide specified amounts of attenuation and/or isolation between circuits. In this section we will consider the lossless matching networks.

When Y_A, Y_B, Y_C or Z_A, Z_B, Z_C are purely reactive, then the network is lossless and $\alpha = 0$. Hence $\theta = j\beta$ and the design equations reduce to

Pi:

$$Y_C = \frac{\sqrt{Y_1\, Y_2}}{j \sin\beta} \tag{6.66}$$

$$Y_B = \frac{Y_2}{j \tan\beta} - Y_C \tag{6.67}$$

$$Y_A = \frac{Y_1}{j \tan\beta} - Y_C \tag{6.68}$$

T:

$$Z_C = \frac{\sqrt{Z_1\, Z_2}}{j \sin\beta} \tag{6.69}$$

$$Z_A = \frac{Z_1}{j \tan\beta} - Z_C \tag{6.70}$$

$$Z_B = \frac{Z_2}{j \tan\beta} - Z_C \tag{6.71}$$

A word of caution on interpreting β when source and/or load admittances are complex. We have seen that when the source and load are resistive, β can be interpreted as the voltage or current phase shift, since voltage and current are in phase at both the input and output. If the source and/or load are complex, then the voltage and current are not in phase. In this situation β has a different interpretation for the Pi- and T-networks. This can be seen by considering Figure 6.28.

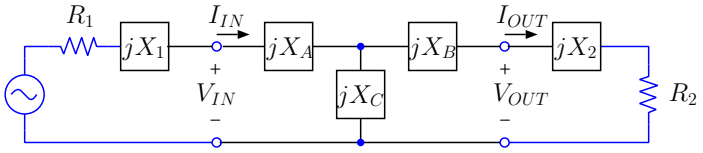


Figure 6.28: T-network with complex source and load

The network would be designed by incorporating X_1 and X_2 into the matching network as shown in Figure 6.29.

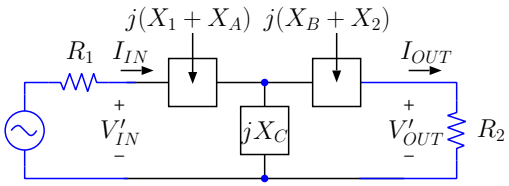


Figure 6.29: Network designed with X_1 and X_2 incorporated into the matching network

The currents I_{IN} and I_{OUT} in Figure 6.29 are the same as those shown in Figure 6.28, but the voltages are not. In fact, V'_{IN} and V'_{OUT} are related to the V_{IN} and V_{OUT} of Figure 6.28 by

$$V'_{OUT} = V_{OUT} \frac{R_2}{R_2 + j\,X_2} \tag{6.72}$$

$$V_{IN} = V'_{IN} \frac{R_1 - jX_1}{R_1} \tag{6.73}$$

Now, β is the phase shift between I_{IN} and I_{OUT} or, equivalently, V'_{IN} and V'_{OUT}. It is not the phase shift between V_{IN} and V_{OUT}, however. The voltage phase shift $(\theta_{V_{IN}} - \theta_{V_{OUT}})$ can be determined as follows:

$$\frac{V_{IN}}{V_{OUT}} = \frac{V'_{IN}}{V'_{OUT}} \; \frac{R_1 - jX_1}{R_2 + jX_2} \; \frac{R_2}{R_1} \tag{6.74}$$

Define

$$\theta_{Z_1} = \tan^{-1} \frac{X_1}{R_1} \tag{6.75}$$

$$\theta_{Z_2} = \tan^{-1} \frac{X_2}{R_2} \tag{6.76}$$

Then

$$\begin{aligned} \theta_{V_{IN}} - \theta_{V_{OUT}} &= \theta'_{V_{IN}} - \theta'_{V_{OUT}} - \theta_{Z_1} - \theta_{Z_2} \\ &= \beta - \theta_{Z_1} - \theta_{Z_2} \end{aligned} \tag{6.77}$$

Thus for the lossless T-network β gives the current phase shift, and the voltage phase shift can be found from Equation 6.77.

Similar considerations lead to the conclusion that β gives the voltage phase shift for the lossless Pi-network. In this case Equation 6.78 yields the current phase shift:

$$\theta_{I_{IN}} - \theta_{I_{OUT}} = \beta - \theta_{Y_1} - \theta_{Y_2} \tag{6.78}$$

6.4.3.1 Example - Design of a lossless Pi-network with specified phase shift

Design a lossless Pi-network to match a load of 300 Ω to a 50 Ω source with V_{OUT} leading V_{IN} by 45° at $\omega = 10^7$ rad/s as in Figure 6.30. We want $\theta_{V_{OUT}} - \theta_{V_{IN}} = 45°$. Hence

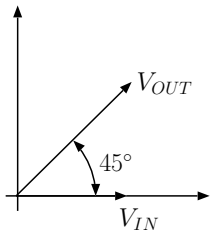


Figure 6.30: Phasor diagram showing V_{OUT} leading V_{IN} by 45°.

$\beta = -45°$

$$\begin{aligned} Y_1 &= \frac{1}{50} &&= 20\,\text{mS} \\ Y_2 &= \frac{1}{300} &&= 3.3\,\text{mS} \\ Y_C &= \frac{\sqrt{20(3.33)}}{j\,\sin(-\pi/4)} &&= +j\,11.54\,\text{mS} \\ Y_B &= \frac{3.33\,\text{mS}}{j\,\tan(-\pi/4)} - j1154\,\text{mS} &&= -j\,8.21\,\text{mS} \\ Y_A &= \frac{20\,\text{mS}}{j\,\tan(-\pi/4)} - j1154\,\text{mS} &&= +j\,8.46\,\text{mS} \end{aligned} \tag{6.79}$$

The final solution is shown in Figure 6.31. The network will have a band-pass transfer

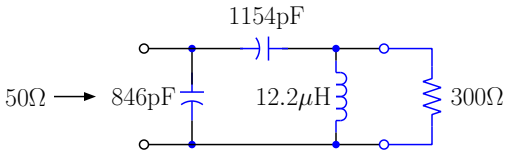


Figure 6.31: Pi-network solution

function characteristic, as the series capacitance and shunt inductor guarantee that the network transfer function will have zero response at DC, whereas the shunt capacitor will provide high frequency attenuation which increases at 6 dB per octave for frequencies well above the design frequency.

Complex source or load impedances can be handled by incorporating the load reactances into the network, as illustrated in the following example.

6.4.3.2 Example - Matching complex load with a specified current phase shift.

At $\omega = 10^7$ rad/s design a lossless matching network to match a load impedance of 150 - j 75 Ω to a generator having an internal impedance of 50 Ω. The output current is to be in

phase with the input current as in Figure 6.32. Start by trying to find a T-network solution:

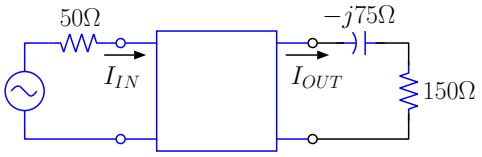


Figure 6.32: Example

$$\begin{aligned} \beta &= \theta_{I_{IN}} - \theta_{I_{OUT}} = 0 \\ \sin\beta &= 0 \\ \tan\beta &= 0 \end{aligned} \tag{6.80}$$

Thus, $Z_A,\ Z_B,\ Z_C \Rightarrow \infty$. This illustrates that 0 current phase shift cannot be obtained with a T-network!

Let us consider a Pi-network. For the Pi-network, β is the voltage phase shift. As noted earlier, β can be written in terms of the current phase shift and the phase angles of the terminating impedances:

$$\beta = \theta_{I_{IN}} - \theta_{I_{OUT}} - \theta_{Z_1} - \theta_{Z_2} = 0 - 0 - 0 - \theta_{Z_2} = -\theta_{Z_2}$$

$$\theta_{Z_2} = \tan^{-1}\frac{-75}{150} = -26.6^\circ \tag{6.81}$$

Thus $\beta = +26.6^\circ$. To proceed, the load is transformed so that absorption can be used as in Figure 6.33. Then

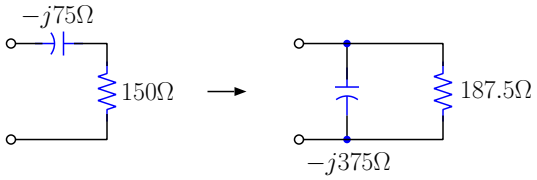


Figure 6.33: Transform load

$$\begin{aligned} Y_C &= \frac{\sqrt{(1/50)(1/187.5)}}{j\ \sin 26.6^\circ} = -j\,23.07\,\text{mS} \\ Y_A &= \frac{1/50}{j\ \tan 26.6^\circ} + j\,23.07\,\text{mS} = -j\,16.9\,\text{mS} \end{aligned} \tag{6.82}$$

$$\frac{-1}{j\,375} + Y_B = \frac{1/187.5}{j\ \tan 26.6^\circ} + j\,23.07\,\text{mS}$$

$$Y_B = j\,9.73\,\text{mS} \tag{6.83}$$

The final result is shown in Figure 6.34.

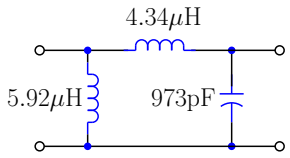


Figure 6.34: Final Result.

This example illustrated how to properly account for phase shifts in complex load impedance. A similar approach would be necessary if the source impedance was complex.

We now turn from the purely reactive 3-element matching networks to a discussion of purely resistive networks.

6.4.4 Resistive Three-element Matching Networks

The purely resistive matching networks are also known as "attenuators," "attenuating pads," or just "pads." Here we assume Y_A, Y_B, Y_C (or Z_A, Z_B, and Z_C) are real (resistances). The resistive matching network can be employed to provide a broadband match between two resistive terminations. Of course, some power loss must be accepted. This type of network is used to build attenuators with specified attenuation. In addition, as we will see, the resistive network can be used to ensure that a particular device sees a well-defined impedance, even when the input of the following device has a variable or unknown impedance.

In this application Y_A, Y_B, Y_C are real (resistive). The terminating impedances are also assumed to be resistive. There will be no phase shift and hence

$$\theta = \alpha \ \text{ (real) nepers} \tag{6.84}$$

As noted previously

$$2\alpha = \ln \frac{P_{IN}}{P_{OUT}} \tag{6.85}$$

Thus, α determines the attenuation of the network. It is useful to relate α to the attenuation in dB:

$$\begin{aligned} \text{Attenuation in dB} &= 10 \log \frac{P_{IN}}{P_{OUT}} \\ &= 20 \log e^{\alpha} \\ &= 8.686\alpha \end{aligned} \tag{6.86}$$

If a network that provides 10 dB of attenuation is required, α would be $10/8.686 = 1.1513$ nepers.

6.4.4.1 Example - Design a 10 dB Pi-type resistive attenuator

Design a 10 dB Pi-type attenuator for use in a system with 75 Ω impedance ($R_1 = R_2 = 75\,\Omega$) as in Figure 6.35.

$$\begin{aligned}
\alpha &= 10/8.686 = 1.1513 \\
Y_1 &= Y_2 = \tfrac{1}{75} = 13.33 \text{ mS} \\
Y_C &= \frac{\sqrt{Y_1 Y_2}}{\sinh\ \alpha} = \frac{13.33 \text{ mS}}{1.423} = 9.370 \text{ mS} \\
Y_A &= Y_B = \frac{13.33 \text{ mS}}{\tanh\ \alpha} - Y_C = 6.922 \text{ mS} \\
Z_A &= Z_B = 144.5\,\Omega \simeq 145\,\Omega \\
Z_C &= 106.7\,\Omega \simeq 107\,\Omega
\end{aligned}$$

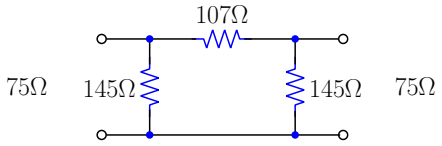


Figure 6.35: 10 dB Pi-type attenuator with 75 Ω impedance

The resistive matching networks/attenuators have the useful characteristic of isolating the impedance found at the input from the impedance that terminates the output. This can be seen by computing the input impedance for the extreme cases of shorted and open-circuit output terminations in the example considered above:

$$\begin{aligned}
\text{shorted output :} \quad Z_{IN} &= 61.6\,\Omega \\
\text{open output :} \quad Z_{IN} &= 92.0\,\Omega
\end{aligned}$$

Clearly, for any resistive termination, the input impedance will be in the range 61.6 to 92.0 Ω. This property can be used to advantage when the input impedance of a particular stage is not well known or is subject to variation. If this stage follows a stage that requires a certain load impedance in order to operate correctly, an attenuator can sometimes be an effective "isolation" stage. For example, for proper operation of certain types of frequency mixers, it is necessary that all of the frequency components at the output of the mixer "see" a well-defined impedance (usually 50 Ω). Typically the frequencies at the output of a mixer may span a very broad range, while the following stage is often a narrow-band IF amplifier. The IF amplifier would often have a well defined input impedance only within its passband. A resistive matching network is sometimes employed between the mixer and the IF amplifier in order to ensure proper termination of the mixer output port.

It is important to note that the attenuation of a resistive matching network will only be equal to the design value if the network is operated between the impedances for which

it was designed. It is important to remember this, since the resistive networks are often found in applications where the terminating impedances are substantially different from the "correct" values. The example cited in the previous paragraph is one such case.

6.4.4.2 Minimum-loss Resistive Matching Networks

In some cases, especially when a broadband impedance match is required, one would like to design a resistive network that has the smallest possible attenuation (α). The problem is then to find the solution that yields the smallest α subject to the constraint that all resistive elements have values greater than or equal to zero. This constraint ensures that the solution corresponds to a passive network.

Consider the T-network in Figure 6.36.

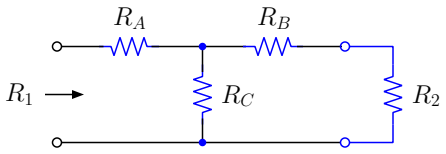


Figure 6.36: T-network

$$R_C = \frac{\sqrt{R_1 R_2}}{\sinh \alpha} \tag{6.87}$$

$$R_B = \frac{R_2}{\tanh \alpha} - \frac{\sqrt{R_1 R_2}}{\sinh \alpha} \tag{6.88}$$

$$R_A = \frac{R_1}{\tanh \alpha} - \frac{\sqrt{R_1 R_2}}{\sinh \alpha} \tag{6.89}$$

The goal is to find the smallest α that gives values for R_A, R_B, R_C which are all ≥ 0. Note from Equation 6.87 that

$$\sinh \alpha = \frac{\sqrt{R_1 R_2}}{R_C} \tag{6.90}$$

Since sinh α increases monotonically as a function of α, as shown in Figure 6.37, minimizing α is equivalent to minimizing sinh α. Clearly then, we should use the largest value of R_C

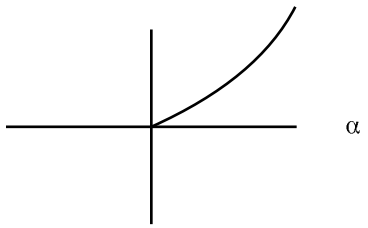


Figure 6.37: sinh α as a function of α

for which $R_B \geq 0$ and $R_A \geq 0$ (see Equation 6.90). Using

$$\cosh \alpha = \sqrt{1 + \sinh^2 \alpha} \tag{6.91}$$

$$= \sqrt{1 + R_1 R_2 / R_C{}^2} \tag{6.92}$$

we find from Equation 6.88 and Equation 6.89

$$R_B = R_2 \sqrt{\frac{R_C^2}{R_1 R_2} + 1} - R_C \tag{6.93}$$

$$R_A = R_1 \sqrt{\frac{R_C^2}{R_1 R_2} + 1} - R_C \tag{6.94}$$

For the moment, assume $R_2 > R_1$. Then, as R_C is increased in order to decrease $\sinh(\alpha)$, R_A will decrease and will become equal to zero when (see Equation 6.94)

$$R_C = \frac{R_1}{\sqrt{1 - R_1/R_2}} \tag{6.95}$$

Be aware that R_B will always be greater than zero if $R_2 > R_1$. So, if $R_2 > R_1$, the minimum attenuation is obtained when R_C is given by Equation 6.95. The corresponding value for α (denoted by α_{min}) is obtained from Equation 6.92 and is

$$\alpha_{min} = \cosh^{-1} \sqrt{R_2/R_1} \tag{6.96}$$

The value for R_B is obtained by using Equation 6.95 in Equation 6.93 to yield

$$R_B = R_2 \sqrt{1 - R_1/R_2} \tag{6.97}$$

Notice that since $R_A = 0$, the minimum-loss resistive network reduces to an L-network with the series arm (R_B) connected to the larger of the two terminating resistances.

The preceding discussion was based on the T-network, although the final result was found to be an L-network. The same result would have been obtained if the starting point had been the Pi-network.

6.4.4.3 Summary of Resistive Minimum-loss Network

The resistive minimum-loss network reduces to an L-network with series arm connected to the larger of the terminating resistances as in Figure 6.38.

$$\text{Loss in dB} = 8.686 \cosh^{-1} \sqrt{R_{big}/R_{small}}$$

$$R_{shunt} = \frac{R_{small}}{\sqrt{1 - R_{small}/R_{big}}} \tag{6.98}$$

$$R_{series} = R_{big} \sqrt{1 - R_{small}/R_{big}}$$

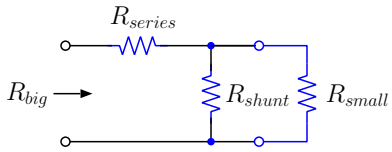


Figure 6.38: Resistive minimum-loss network

6.5 References

1. Smith, Jack, *Modern Communication Circuits*, McGraw-Hill, 1986.

2. Terman, Frederick Emmons, *Radio Engineers Handbook*, McGraw Hill, 1943.

6.6 Homework Problems

1. You are given a "black box" with two output terminals (a "1-port"). You play around with the box for awhile and make the following observations:

   (a) The output voltage from the box is sinusoidal.

   (b) The **peak magnitude** of the open circuit voltage at the output of the box is found to be 5 V.

   (c) You connect a 50 Ω resistor across the terminals and find that the peak magnitude of the voltage across the resistor is 2.795 V.

   (d) You short the output of the box and find the **peak magnitude** of the short circuit current is 100 mA.

   Find the power available from the source. Express your result in dBm.

2. Consider a source with impedance $Z_S = R_S + jX_S$ and suppose that the source drives a load that is purely resistive. Denote the load resistance by R_L.

   (a) What load resistance should be used if the goal is to maximize the power delivered to the load?

   (b) Find an expression for the power delivered to the load resistance found in part 2a. Express your result in terms of the source parameters only, i.e., P_{avs}, R_S and X_S.

3. Find the lossless network having the *minimum number of elements* to match a source impedance of 20-j60 Ω to a load impedance having an equivalent circuit of 100 Ω of resistance shunted by 50 Ω of capacitive reactance. Draw the network and label all parts.

4. Consider the design of a lossless L-network to match the source and load shown in Figure 6.39. All resistances and reactances are in ohms, and the current source magnitude is the peak value.

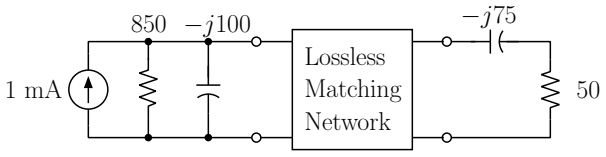


Figure 6.39: Complex source and load to be matched with a lossless L-network.

   (a) Find the power available from the source. Express your result in dBm.

   (b) How much power would be delivered to the load if a lossless matching network was not used, i.e., if the load is connected directly to the source? Express your result in dBm.

(c) There are four possible solutions for the matching network if an L-network is used. Find all four. Sketch all solutions and indicate whether the elements are inductors or capacitors.

(d) Verify two of your designs by plotting the path from the load to the source on a Smith Chart.

5. Consider a source with impedance $Z_S = R_S + jX_S$ and a load that is purely resistive. Denote the load resistance by R_L. Suppose that $R_S < R_L$. Under what condition(s) are there more than 2 lossless L-networks that conjugately match the source to the load? Derive a simple inequality that can be checked to determine whether there are 2 or 4 solutions.

6. The source and load shown in Figure 6.40 can be matched with an L-network that consists of two capacitors. All resistances and reactances are given in ohms.

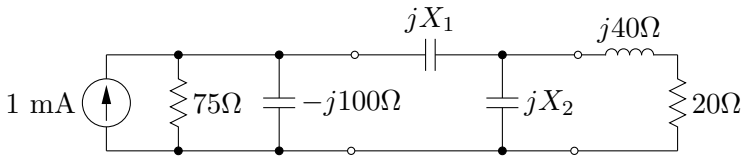


Figure 6.40: Source and load matched with 2-capacitor L-network.

Find X_1 and X_2 that will cause all of the available source power to be delivered to the load. Both X_1 and X_2 must be <0.

7. Design a matching network to match the source and load shown in Figure 6.41. Use a T-network such that I_{out} lags I_{in} by 60° at $\omega = 10^7$. What is the voltage phase shift?

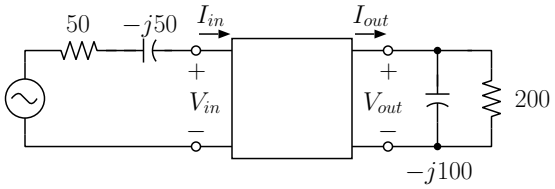


Figure 6.41: Source and load for lossless T-network.

8. The circuit in Figure 6.42 is matched at $\omega = 10^8\, rad\, s^{-1}$. Redesign for a match with the second harmonic attenuated in the shunt arm connected to the 100 Ω resistor and the third harmonic attenuated in the shunt arm connected to the 400 Ω resistor.

9. Consider the design of a band-pass T-type matching network with specified Q. This can be approached by thinking of the T-net as two back-to-back L-networks. One of the L-networks has a high-pass topology and the other has low-pass topology.

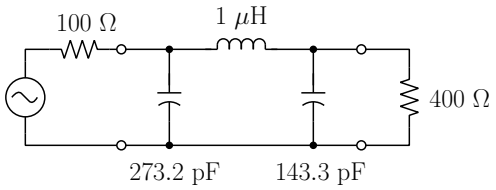


Figure 6.42: Circuit matched at $\omega = 10^8\ s^{-1}$.

Find a band-pass T-network that will match a 50 Ω load to a 10 Ω source impedance. Find the solution with a series inductor connected to the 10 Ω termination and a capacitor connected to the 50 Ω termination. Choose the overall Q of the network to be approximately equal to 10, i.e., choose the virtual resistance, R_v, such that the L-net with the highest Q has Q=10. Draw the T-network and label all components. Note: You should combine the two shunt elements into one single element for this problem.

10. Consider in Figure 6.43 the design of a T-network that transforms a 50 Ω load impedance to 5 Ω:

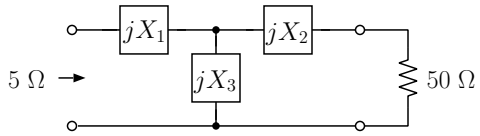


Figure 6.43: T-network transforming 50 Ω load impedance to 5 Ω.

Design a T-network with a Q of 6. Find the solution with $X_1 > 0$ and $X_2 < 0$. Specify X_1, X_2 and X_3 to the nearest ohm.

11. Consider the design of a T-network that transforms a 50 Ω load impedance to 10 Ω. Suppose that the reactance of the shunt element of the T-network is fixed at +j200 Ω as shown in Figure 6.44:

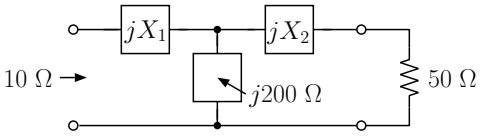


Figure 6.44: T-network transforming 50 Ω load impedance to 10 Ω.

Find two sets of values (X_1, X_2) which will give a match.

12. Design a T-network (as shown in Figure 6.45) that does an impedance inversion, i.e., that gives the following mapping between Z_L and Z_{IN} where R_0 is a positive real constant:

$$Z_{IN} = \frac{R_0^2}{Z_L} \tag{6.99}$$

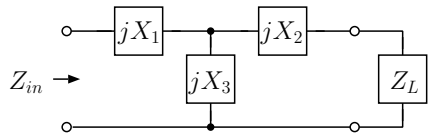


Figure 6.45: T-network providing impedance inversion.

Find expressions for X_1, X_2 and X_3. You may assume that $X_1 > 0$.

13. Consider the design of the matching network in Figure 6.46. Find the reactance X_s and the turns ratio of the ideal transformer, n. The ideal transformer is characterized by the relations:

$$V_2 = nV_1 \quad \text{and} \quad I_2 = -I_1/n \tag{6.100}$$

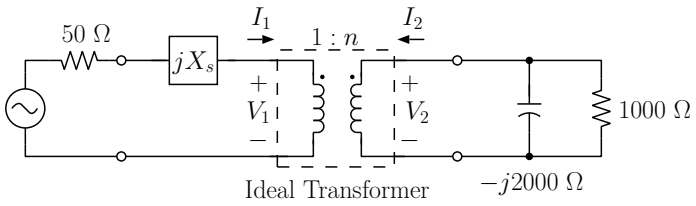


Figure 6.46: Matching network to find X_s and N

14. Suppose that you build an L-type matching network using a variable inductor and capacitor as shown in Figure 6.47. This network is to be used to match the impedance of a load, Z_L, to the output of a generator that has a 50 Ω impedance. At some frequency you know that the inductor's reactance can be adjusted to any value in the range from +j10 to +j100 Ω. At the same frequency the capacitor's reactance can be adjusted to any value in the range from -j10 to -j100 Ω. On a Smith Chart show the region corresponding to all possible values of load impedance, Z_L, that can be matched to 50 Ω using this network. Normalize your Smith Chart to 50 Ω. Shade or color the region of the chart that corresponds to the "matchable" load impedances.

15. Consider the network shown in Figure 6.48. The resistor, capacitor, and inductor are variable with $25\,\Omega \le R \le 50\,\Omega$, $25\,\Omega \le X_L \le 50\,\Omega$, and $10\,\Omega \le X_C \le 40\,\Omega$. On a

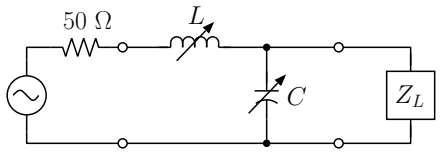


Figure 6.47: L-network with variable inductor and capacitor.

Smith Chart show the region corresponding to all possible values of input admittance, Y_{in}. Normalize your Smith Chart to 50 Ω.

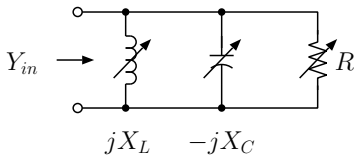


Figure 6.48: Network with variable resistor, capacitor and inductor.

16. Design a resistive T-type attenuator that results in 10 dB attenuation when used with 50 Ω source and load impedances.
    (a) Consider all possible resistive loads and find upper and lower bounds on the input resistance of the attenuator.
    (b) Now design a resistive Pi-type attenuator for the same parameters as the T-type attenuator.
    (c) Repeat part 16a for the Pi-type attenuator.
    (d) Suppose the Pi-type attenuator is used with a 600 Ω load. Find the attenuation in dB where attenuation is defined to be 10 log (P_{in}/P_{out}).
17. An 8 dB attenuator that is designed to match a 75 Ω source to a 300 Ω load is used between a 75 Ω source and a 10 Ω load. The source has available power of 10 dBm. Find the power delivered to the 10 Ω load. Express your result in dBm.
18. Consider a resistive attenuator/matching-network that is designed to work with 50 Ω source and load resistances, as in Figure 6.49. When connected to 50 Ω terminations, the attenuator is designed to give an attenuation of 10 dB, i.e., $P_{in}/P_{out} = 10$, and also to provide a 50 Ω match at both input and output. Now, suppose that the attenuator is used with an arbitrary load resistance, R_L (the source impedance is still 50 Ω). Denote the power available from the 50 Ω source by P_{avs} and find an expression for the power delivered to the load resistor, R_L. Your result should be in terms of P_{avs} and R_L only (and numerical constants).
19. Design a minimum-loss resistive matching network that will match a 300 Ω source to a 50 Ω load. What is the attenuation in dB?

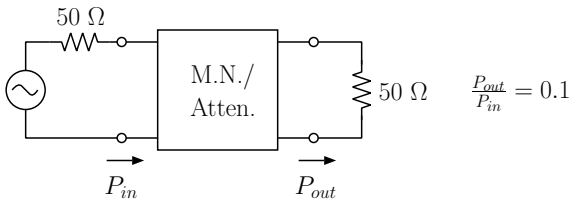


Figure 6.49: Resistive attenuator/matching-network.

20. Suppose that you have a lossless matching network of unknown type that transforms 50 Ω (resistive) to 450 Ω (resistive) and you need to couple a 50 Ω source to a resistive load, R_L, which may take on any positive, real, value. You have the choice of connecting the 50 Ω source directly to R_L, or using the matching network between the source and R_L (with the 50 Ω side of the matching network connected to the source). For what values of R_L would more power be delivered to R_L with the matching network in the system?

21. Design a lossless L-network that will match a 5 Ω source to a 200 Ω load. Use an L-network with high-pass topology.

    (a) Draw the network and indicate the reactances of the series and shunt elements.

    (b) Now, suppose that you build the network that you designed using a lossy inductor that has Q_L=32. Because the network contains a lossy element, the power that is actually delivered to the 200 Ω resistor will be smaller than the power available from the source. Find the loss of the network. The loss is defined as the difference (in decibels) between P_{avs} and the power that is actually delivered to the 200 Ω load.

22. Design the lossless L-network having a lowpass topology and with the series arm connected to the source, as shown in Figure 6.50.

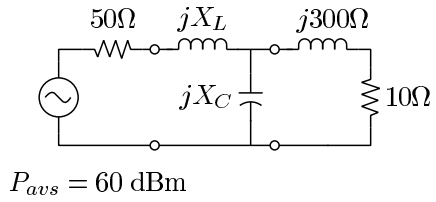


Figure 6.50: Lossless L-network with lowpass topology

    (a) Specify X_C and X_L.

    (b) Suppose the matching network you just designed is used and the imaginary part of the load impedance increases by 10%, i.e., suppose the load impedance changes to

$10 + j\ 330\ \Omega$. Determine the change in the power delivered to the load. Express your result in dB and be sure to correctly indicate the sign of the change.

23. A 50 Ω source has $P_{avs} = 4\,\text{mW}$. It is necessary to couple the source to a load with impedance $Z_L = 2 + j20\,\Omega$.

    (a) Find the power delivered to the load when the load is connected directly to the source. Express your answer in dBm.

    (b) There are 4 lossless LC L-networks that will match this source and load. Find the solution that has a capacitive series arm and the shunt arm connected across the load. You will receive full credit only if you specify the L-network having the specified topology.

24. Consider the source and load shown in the Figure.

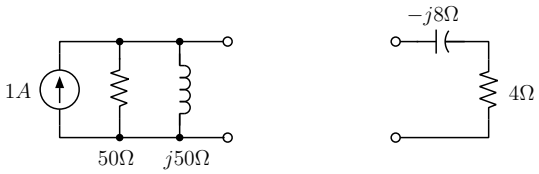


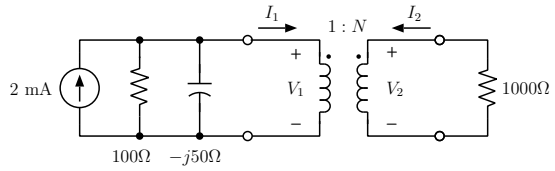
Figure 6.51: Source and load for problem 24.

    (a) Find the power available from the source. Express your answer in dBm.

    (b) Match this source and load using a lossless L-network. Any valid solution will be accepted.

    (c) Now suppose that you can only use a single lossless inductor or capacitor, either in series or in shunt, to couple the source to the load. If the goal is to maximize the power delivered to the load under this constraint, find the best possible solution.

25. A source is coupled to a resistive load, $R_L = 1000\,\Omega$, through an ideal transformer. The specified source current is the peak value. The ideal transformer is described by the relations: $V_2 = NV_1$ and $I_2 = -I_1/N$.



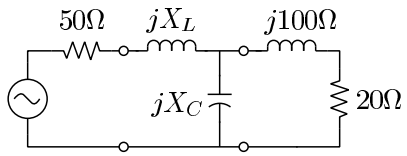
    (a) Find the power available from the source. Express your result in dBm.

    (b) Assuming that the turns ratio, N, is a positive real number find the value of N that maximizes the power delivered to the load.

(c) Assuming that N is set equal to the optimum value found in part (b.), determine the power delivered to the load. Express your result in dBm.

(d) Suppose the transformer is removed from the system, i.e. suppose that the $1000\,\Omega$ load is connected directly to the source. Find the power that would be delivered to the load. Express your result in dBm.

26. Design the lossless L-network with a low-pass topology having the series arm connected to the source, as shown below.



$P_{avs} = 10$ dBm

(a) Specify X_L and X_C.

(b) Determine the power that would be delivered to the $(20 + j100)\Omega$ load if the source was connected directly to the load. Express your result in dBm.

(c) Suppose the L-network designed in part a. is replaced with a new network consisting of a single capacitor. Sketch the single-capacitor network that will maximize the power delivered to the load. Specify the reactance of the capacitor and the power that would be delivered to the load (in dBm).

27. Consider the problem of coupling a $200\,\Omega$ source with $P_{avs} = 0$ dBm to the load $Z_L = 10 + j100\,\Omega$.

(a) Design a passive, lossless L-network that matches the source to the load and consists of 2 capacitors (no inductors may be used). Find the solution with the shunt element of the L-network connected across the load. Sketch the system, including the source, the matching network, and the load, and label each capacitor with its impedance.

(b) An ideal, lossless transformer with turns ratio N:1 is used between the given source and load. The turns ratio is a positive real number, and N>1. Find N that results in the largest possible power delivered to the load. Include a sketch showing the source, the transformer, and the load. Be sure to indicate the correct orientation of the transformer (i.e. specify which side of the transformer corresponds to the larger number of turns "N").

(c) Find the power delivered to the load if the transformer specified in part b. is used between the source and load. Express your result in dBm.

28. An ideal transformer can be used to reduce the mismatch loss (increase the mismatch factor) between an arbitrary source impedance $Z_S = R_S + jX_S$ $(R_S > 0)$ and a pure resistance R_L as shown in the Figure. Suppose that the turns ratio, N, of the transformer can be adjusted to any positive real number $(N > 0)$.

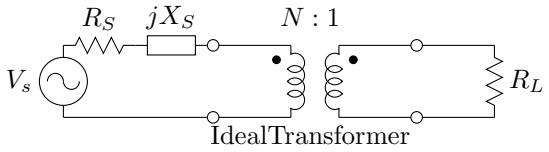


Figure 6.52: Ideal transformer matching complex source to real load.

(a) For given Z_S and R_L, there is an optimum turns ratio, N_{opt}, that maximizes the power delivered to the load. Find an expression for N_{opt}.

(b) Assuming that the optimum turns ratio is used, the resulting mismatch factor ($MF = \frac{P_{out}}{P_{avs}}$) depends only on the phase angle of the source impedance, i.e. if $Z_S = |Z_S|e^{j\theta}$, the mismatch factor can be written in terms of θ. Find such an expression for the mismatch factor.

Chapter 7

Introduction to 2-port Parameters

7.1 Introduction

An *n-port* network has 2n terminals that are grouped in pairs to form n *ports*. The i'th port has a voltage, V_i, between its terminals and a current I_i flowing in to one of the terminals and out of the other terminal. In this chapter we will be concerned with 2-port networks under sinusoidal steady-state excitation. Figure 7.1 summarizes the voltage and current conventions that will be used. Voltages and currents are assumed to be phasor quantities.

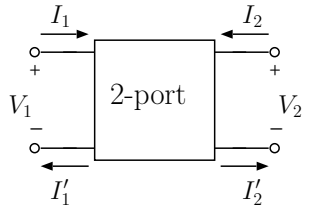


Figure 7.1: Voltage and current conventions

Notice that a port has the property that the value of the current flowing in to the port through the upper terminal is always equal to the current flowing out of that port through the lower terminal, i.e. refer to Figure 7.1 and note that this property means that $I_1 = I_1'$ and $I_2 = I_2'$. If two terminals are chosen whose currents do not satisfy this property, then they cannot be said to comprise a single port - instead, they must be considered to be terminals associated with two different ports.

The discussion in this chapter will concentrate on characterizing the properties of *linear*, *time-invariant* (LTI) 2-ports. For our purposes, any 2-port comprised of linear network elements (i.e. including linear resistors, linear inductors, linear capacitors, linear transformers, linear controlled sources, linear transmission lines, and independent sources) is a *linear* network. *Time-invariance* simply means that the parameters of the network elements do not change with time.

Only four complex numbers are needed to completely characterize the properties of a LTI 2-port network at one frequency. The four complex parameters can be interpreted

in terms of input and output impedances (or admittances) and forward and reverse gains when the 2-port is terminated with specific source and load impedances. The standard, or reference, source and load impedances are constants for a given parameter set. The properties of the 2-port when it is terminated with the reference impedances can be used to predict the properties of the 2-port in a system with arbitrary terminating impedances. In particular, given the four complex numbers that comprise a specific 2-port parameter set, it is possible to determine, for any combination of source and load terminations, the input and output impedances, voltage and current gains, and power gain. It is also possible to determine whether or not the 2-port can oscillate with passive source and load terminations and, if so, the region of the impedance plane (or Smith Chart) containing the source and load terminations that support oscillation can be identified.

There are various ways to define the four complex numbers that describe a 2-port. The different possibilities arise from considering two of the four variables (V_1, I_1, V_2, I_2) to be *dependent* on the other two (*independent*) variables. There are six possible choices for the pair of dependent variables, leading to six possible parameter sets defined in terms of voltage and current variables. The four most commonly used parameter sets are the Y, Z, h, and ABCD parameters. In the next Chapter we will consider two more parameter sets (S and T parameters) that are defined using variables that are linear combinations of the voltage and current at a particular port. In any case, once one set of parameters is known it can be translated into any of the other parameter sets without making additional measurements.

7.1.1 Y-parameters

Consider the *Y-parameters*, or *admittance parameters*:

$$\begin{aligned} I_1 &= Y_{11}V_1 + Y_{12}V_2 \\ I_2 &= Y_{21}V_1 + Y_{22}V_2 \end{aligned} \tag{7.1}$$

The voltages are the independent variables and the currents are dependent variables. Measurement of the Y-parameters involves forcing one of the independent variables to be zero, i.e., if V_2 is forced to zero by placing a zero impedance load across the output terminals, then:

$$\begin{aligned} Y_{11} &= \left.\frac{I_1}{V_1}\right|_{V_2=0} \quad \text{Input admittance with output shorted} \\ Y_{21} &= \left.\frac{I_2}{V_1}\right|_{V_2=0} \quad \text{Forward transfer admittance, output shorted} \end{aligned} \tag{7.2}$$

Likewise, if V_1 is forced to zero by shorting the input, Y_{22} and Y_{12} can be measured using:

$$\begin{aligned} Y_{22} &= \left.\frac{I_2}{V_2}\right|_{V_1=0} \quad \text{Output admittance with input shorted} \\ Y_{12} &= \left.\frac{I_1}{V_2}\right|_{V_1=0} \quad \text{Reverse transfer admittance, input shorted} \end{aligned} \tag{7.3}$$

All of the Y-parameters have units of admittance. Y-parameters are determined experimentally by measuring input and transfer admittances with one port driven and the other port terminated in a short circuit. Therefore the standard, or reference, termination for the Y-parameter set is a short circuit.

7.1.2 Z-parameters

Z-parameters or impedance parameters were used in chapter 5 to characterize the properties of transformers. Z-parameters are defined such that the currents are the independent variables and the voltages are dependent, i.e.,

$$\begin{aligned} V_1 &= Z_{11}I_1 + Z_{12}I_2 \\ V_2 &= Z_{21}I_1 + Z_{22}I_2 \end{aligned} \tag{7.4}$$

The Z-parameters can be interpreted as follows:

$$\begin{aligned} Z_{11} &= \left.\frac{V_1}{I_1}\right|_{I_2=0} && \text{Input impedance, output open circuited} \\ Z_{21} &= \left.\frac{V_2}{I_1}\right|_{I_2=0} && \text{Forward transfer impedance, output open circuited} \\ Z_{22} &= \left.\frac{V_2}{I_2}\right|_{I_1=0} && \text{Output impedance, input open circuited} \\ Z_{12} &= \left.\frac{V_1}{I_2}\right|_{I_1=0} && \text{Reverse transfer impedance, input open circuited} \end{aligned} \tag{7.5}$$

Z-parameters have units of impedance, and measurement of Z-parameters involves determining input, output, and transfer impedances with one of the ports driven and the other open-circuited. Thus, the reference impedance for the Z-parameters is an open circuit.

7.1.3 Hybrid (h) parameters

$$\begin{aligned} V_1 &= h_{11}I_1 + h_{12}V_2 \\ I_2 &= h_{21}I_1 + h_{22}V_2 \end{aligned} \tag{7.6}$$

Hybrid parameters or h-parameters use both a short circuit and an open circuit as reference terminations. The reference termination is an open circuit for the input port and a short circuit for the output port. This is easily recognized by considering h_{11} and h_{22}:

$$h_{11} = \left.\frac{V_1}{I_1}\right|_{V_2=0} \quad \text{Input impedance with output short circuited} \tag{7.7}$$

$$h_{22} = \left.\frac{I_2}{V_2}\right|_{I_1=0} \quad \text{Output admittance with input open circuited} \tag{7.8}$$

Note that the parameters h_{11} and h_{22} have different units (impedance and admittance, respectively). This is the origin of the "hybrid" designation.

A common characteristic of the Y-, Z-, and h-parameter sets is that the independent variables are forced to zero when the 2-port is terminated in a standard, or reference, impedance. For example, the independent variables in the Y-parameter representation are the voltages at the input and output ports. To force one of the independent parameters to zero, it is necessary to terminate one of the ports with a short circuit. The independent variables in the Z-parameter representation are the input and output currents. Since input

or output currents are forced to zero when a port is terminated with an open circuit, the reference impedance for the Z-parameter set is said to be an open circuit. The reference impedance for the h-parameter set is an open circuit for the input port and a short circuit for the output port. Thus the choice of independent variables determines the reference impedance for the parameter set.

7.1.4 ABCD-parameters

ABCD-parameters are defined such that the port 1 variables depend on the port 2 variables:

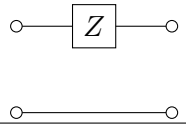
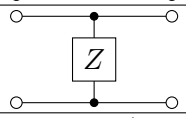
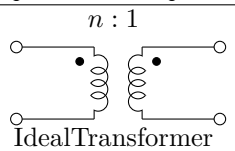
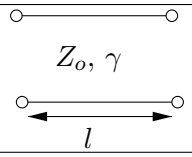
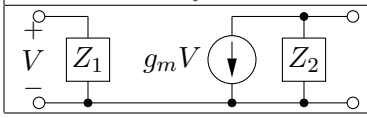
$$\begin{aligned} V_1 &= AV_2 - BI_2 \\ I_1 &= CV_2 - DI_2 \end{aligned} \tag{7.9}$$

The ABCD-parameters can be interpreted or calculated as follows:

$$\begin{aligned} A &= \left(\frac{V_2}{V_1}\right)^{-1}\Big|_{I_2=0} && \text{Inverse open circuit forward voltage gain.} \\ B &= -\left(\frac{I_2}{V_1}\right)^{-1}\Big|_{V_2=0} && \text{Neg. inverse short circuit forward transadmittance.} \\ C &= \left(\frac{V_2}{I_1}\right)^{-1}\Big|_{I_2=0} && \text{Inverse open circuit foward transimpedance.} \\ D &= -\left(\frac{I_2}{I_1}\right)^{-1}\Big|_{V_2=0} && \text{Neg. inverse short circuit forward current gain.} \end{aligned} \tag{7.10}$$

The ABCD parameter representation differs from the Y-, Z-, and h-parameter sets in that the independent variables are both associated with the output port. Thus, it is not possible to define a fixed reference impedance termination for the input and output ports that will force the independent variables to zero. Inspection of equations 7.10 shows that the output port is terminated with an open for calculation of A and C and with a short circuit for calculation of B and D. In any case, this parameter set is very useful because the ABCD matrix for a cascade of 2-ports is the matrix product of the individual ABCD matrices, as shown in the next section.

7.1.5 2-port parameters for some common 2-port networks

2-port	[Z]	[Y]	[ABCD]
	NA	$\frac{1}{Z}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & Z \\ 0 & 1 \end{bmatrix}$
	$Z\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	NA	$\begin{bmatrix} 1 & 0 \\ \frac{1}{Z} & 1 \end{bmatrix}$
	NA	NA	$\begin{bmatrix} n & 0 \\ 0 & \frac{1}{n} \end{bmatrix}$
	$Z_o\begin{bmatrix} \frac{1}{\tanh\gamma l} & \frac{1}{\sinh\gamma l} \\ \frac{1}{\sinh\gamma l} & \frac{1}{\tanh\gamma l} \end{bmatrix}$	$\frac{1}{Z_o}\begin{bmatrix} \frac{1}{\tanh\gamma l} & \frac{-1}{\sinh\gamma l} \\ \frac{-1}{\sinh\gamma l} & \frac{1}{\tanh\gamma l} \end{bmatrix}$	$\begin{bmatrix} \cosh\gamma l & Z_o\sinh\gamma l \\ \frac{1}{Z_o}\sinh\gamma l & \cosh\gamma l \end{bmatrix}$
	$\begin{bmatrix} Z_1 & 0 \\ -g_m Z_1 Z_2 & Z_2 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{Z_1} & 0 \\ g_m & \frac{1}{Z_2} \end{bmatrix}$	$-\frac{1}{g_m}\begin{bmatrix} \frac{1}{Z_2} & 1 \\ \frac{1}{Z_1 Z_2} & \frac{1}{Z_1} \end{bmatrix}$

7.2 Special types of 2-ports and their matrix properties

7.2.1 Reciprocal 2-ports

A 2-port network that does not contain any sources and is constructed only from any combination of resistors, inductors, capacitors, transformers and transmission lines will obey the principle of *reciprocity*, which states that the current (voltage) response at one port due to a voltage (current) excitation at another port is independent of which port is excited and which port the response is measured at. Figure 7.2 illustrates this principle for the case

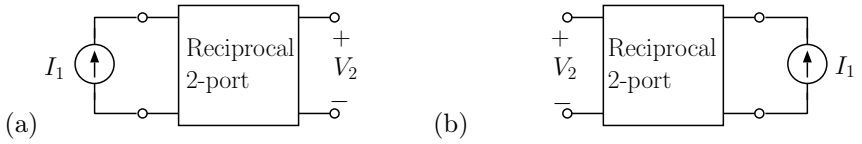


Figure 7.2: If the 2-port is reciprocal, voltage response at port 2 due to current I_1 at port 1 is the same as voltage response at port 1 due current I_1 at port 2.

of the voltage response due to a current excitation. Since the voltage response at port i due to a current excitation at port j is Y_{ij}, reciprocal 2-ports satisfy $Y_{12} = Y_{21}$. Note that Y_{11} is not necessarily equal to Y_{22} in a reciprocal network. If the network is *reciprocal* and *symmetric*, then $Y_{12} = Y_{21}$ and $Y_{11} = Y_{22}$.

Figure 7.3 illustrates the reciprocity principle for the case of the current response due to

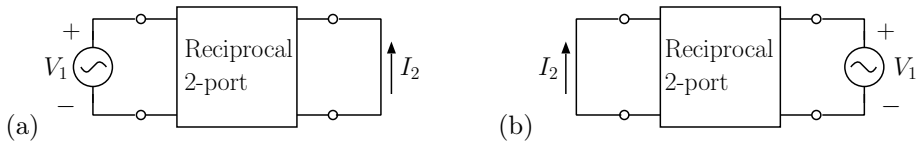


Figure 7.3: If the 2-port is reciprocal, current response at port 2 due to voltage V_1 at port 1 is the same as current response at port 1 due to voltage V_1 at port 2.

a voltage excitation. A reciprocal 2-port will have $Z_{12} = Z_{21}$. A 2-port that is reciprocal and symmetric will have $Z_{12} = Z_{21}$ and $Z_{11} = Z_{22}$.

It is left as an exercise to show that the ABCD matrix of a reciprocal network will satisfy $AD - BC = 1$. In other words, the determinant of the ABCD matrix is unity for a reciprocal 2-port.

The lossless and lossy matching networks discussed in Chapter 6 are all examples of reciprocal networks. Since two of the network parameters are the same, a reciprocal network can be fully characterized (at each frequency) by 3 complex network parameters.

7.2.2 Reciprocal lossless 2-ports

A lossless 2-port contains no dissipative elements, and hence the time-averaged real power absorbed by the network is zero. A 2-port network constructed only of lossless transmission lines, transformers, and lossless L's and C's will be reciprocal and lossless. The constraint that the power absorbed by the network is zero can be written as

$$\frac{1}{2}\Re(V_1 I_1^*) + \frac{1}{2}\Re(V_2 I_2^*) = 0, \tag{7.11}$$

where $\Re(\ )$ takes the real part of its argument. Using the Z parameters, equation 7.11 can be written as

$$\Re(Z_{11}|I_1|^2 + Z_{12} I_1^* I_2 + Z_{21} I_1 I_2^* + Z_{22}|I_{22}|^2) = 0.$$

If the network is reciprocal, then $Z_{12} = Z_{21}$ so the constraint is

$$R_{11}|I_1|^2 + 2R_{21}\Re(I_1 I_2^*) + R_{22}|I_2|^2 = 0, \tag{7.12}$$

where the $R_{ij} = \Re(Z_{ij})$. Since equation 7.12 must be satisfied for any set of excitations, (I_1, I_2), we conclude that $R_{ij} = 0$ for a lossless, reciprocal 2-port. In other words, the Z_{ij} are purely imaginary. Similar reasoning can be used to show that the Y_{ij} are purely imaginary as well.

7.3 Parallel, series, cascade connections of 2-port networks

Any 2-port network can be viewed as being constructed from simpler 2-port networks whose ports are connected in various ways. Analysis of complex 2-port networks can be greatly simplified if the network is decomposed into a number of simpler interconnected 2-port

networks whose 2-port parameters are easily found. In the following sections, the parallel, series, and cascade connections are considered. Other possibilities, such as parallel-series (input ports connected in parallel and output ports in series) and series-parallel are left as exercises.

In all of the following discussions related to interconnection of 2-ports, it is assumed that the interconnection does not upset the basic property that says that the current flowing into one terminal associated with any port must flow out of the other terminal, i.e. referring back to Figure 7.1 we assume that the relationships $I_1 = I_1'$ and $I_2 = I_2'$ are valid for each of the interconnected 2-ports.

7.3.1 2-ports connected in parallel

Figure 7.4 shows a 2-port created by interconnecting two 2-ports with their input ports and output ports connected in parallel. The Y-parameters of the constituent 2-ports can be combined to obtain the Y-parameters of their parallel combination, as shown below.

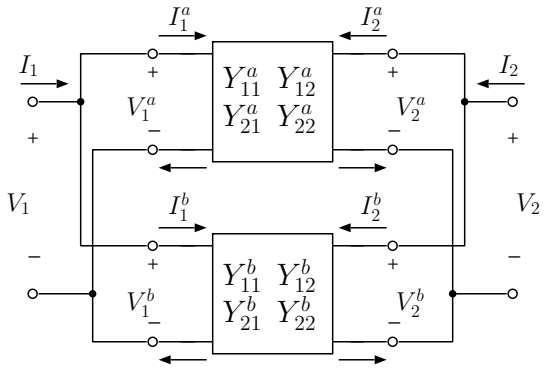


Figure 7.4: Two 2-ports interconnected such that the input and output ports are in parallel.

The parallel connection results in the following relationships:

$$V_1 = V_1^a = V_1^b, \;\; V_2 = V_2^a = V_2^b \tag{7.13}$$

$$I_1 = I_1^a + I_1^b, \;\; I_2 = I_2^a + I_2^b. \tag{7.14}$$

In vector notation, e.g.

$$\mathbf{V} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}, \mathbf{I} = \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix},$$

equations 7.13 and 7.14 can be written as:

$$\mathbf{V} = \mathbf{V}^a = \mathbf{V}^b \tag{7.15}$$

$$\mathbf{I} = \mathbf{I}^a + \mathbf{I}^b. \tag{7.16}$$

The voltage and current vectors associated with the individual 2-ports must satisfy

$$\mathbf{I}^a = \mathbf{Y}^a\mathbf{V}^a \tag{7.17}$$

$$\mathbf{I}^b = \mathbf{Y}^b\mathbf{V}^b. \tag{7.18}$$

Equations 7.15, 7.17, and 7.18 can be used in equation 7.16 to show that

$$\mathbf{I} = \{\mathbf{Y}^a + \mathbf{Y}^b\}\mathbf{V}. \tag{7.19}$$

Thus, the Y-parameter matrix for the parallel combination of 2-ports is the sum of the constituent Y-parameter matrices.

7.3.2 2-ports connected in series

Figure 7.5 shows a 2-port created by interconnecting two 2-ports with their input ports and output ports connected in series. The Z-parameters of the constituent 2-ports can be combined to obtain the Z-parameters of their series combination, as shown in Figure 7.5.

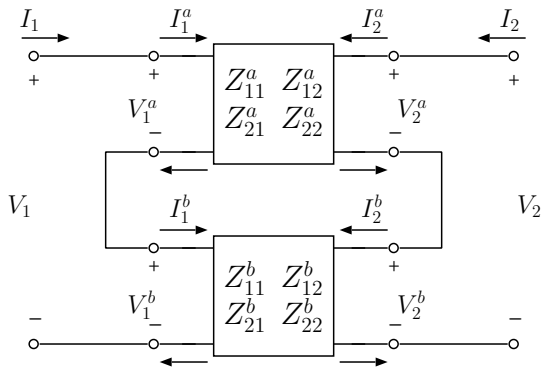


Figure 7.5: Two 2-ports interconnected such that the input and output ports are in series.

The series connection results in the relationships:

$$I_1 = I_1^a = I_1^b, \;\; I_2 = I_2^a = I_2^b$$

$$V_1 = V_1^a + V_1^b, \;\; V_2 = V_2^a + V_2^b.$$

In vector notation the voltage and current vectors must satisfy

$$\mathbf{I} = \mathbf{I}^a = \mathbf{I}^b \tag{7.20}$$

$$\mathbf{V} = \mathbf{V}^a + \mathbf{V}^b. \tag{7.21}$$

The voltage and current vectors for the individual 2-ports must satisfy:

$$\mathbf{V}^a = \mathbf{Z}^a\mathbf{I}^a \tag{7.22}$$

$$\mathbf{V}^b = \mathbf{Z}^b\mathbf{I}^b \tag{7.23}$$

Equations 7.20, 7.22, and 7.23 can be used in equation 7.22 to show:

$$\mathbf{V} = \{\mathbf{Z}^a + \mathbf{Z}^b\}\mathbf{I}. \tag{7.24}$$

Thus, the Z-parameter matrix for the series combination of 2-ports is the sum of the constituent Z-parameter matrices.

7.3.3 Cascaded 2-ports

Figure 7.6 shows two 2-ports connected in cascade, i.e. the output of the first 2-port drives the input of the second 2-port.

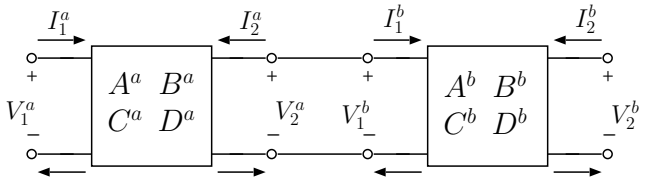


Figure 7.6: Two 2-ports in cascade.

The cascade connection results in the following relationships:

$$V_2^a = V_1^b, \;\; I_2^a = -I_1^b \tag{7.25}$$

Define the input vector $\mathbf{IN} = \begin{bmatrix} V_1 \\ I_1 \end{bmatrix}$, output vector $\mathbf{OUT} = \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix}$, and chain parameter matrix $\mathbf{ABCD} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$. Then the relationships given in equation 7.25 can be written as:

$$\mathbf{OUT}^a = \mathbf{IN}^b \tag{7.26}$$

The following equations are satisfied by the individual 2-ports:

$$\mathbf{IN}^a = \mathbf{ABCD}^a\mathbf{OUT}^a \tag{7.27}$$

$$\mathbf{IN}^b = \mathbf{ABCD}^b\mathbf{OUT}^b \tag{7.28}$$

Using equation 7.28 in equation 7.26, and then using the result in equation 7.27 we find

$$\mathbf{IN}^a = \mathbf{ABCD}^a\mathbf{ABCD}^b\mathbf{OUT}^b. \tag{7.29}$$

Thus, the input and output vectors of the overall 2-port are related by the matrix product of the constituent **ABCD** matrices.

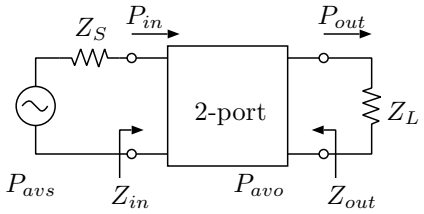


Figure 7.7: A system consisting of a source, 2-port, and a load. The source and load impedances, Z_S and Z_L, may be complex. P_{avs} is the power available from the source and P_{avo} is the power available from the output of the 2-port. P_{in} is the power delivered to the input of the 2-port and P_{out} is the power delivered to the load.

7.4 Power Gain Definitions

One very useful application for the 2-port parameters defined earlier is the calculation of the input and output impedances and the power gains of a 2-port when terminated with arbitrary source and load terminations, Z_S and Z_L. Consider a system consisting of a 2-port driven by a source with impedance Z_S and terminated with a load having impedance Z_L as shown in Figure 7.7.

In general, the input impedance of the 2-port will depend on the 2-port's network parameters and the load impedance, Z_L. The output impedance of the 2-port will depend on the network parameters and the source impedance, Z_S.

The input impedance may not be equal to the conjugate of the source impedance, so the power that is delivered to the input of the 2-port, P_{in}, may be only a fraction of the power that is available from the source. Likewise, the output port can be considered to be a source with impedance equal to Z_{out} and with available power P_{avo}. In general, the output impedance of the 2-port will not be equal to the conjugate of the load impedance, so the power that is delivered to the load, P_{out}, will be some fraction of the power that is available from the 2-port. The relationships between the various powers can be expressed in terms of several different power gains, which are defined next.

1. **Operating Power Gain** (or just "power gain")

$$G \equiv \frac{\text{Power delivered to load}}{\text{Power delivered to } 2-\text{port}} = \frac{P_{out}}{P_{in}} \tag{7.30}$$

   The operating power gain will depend only on the load impedance, Z_L, and on the 2-port network parameters.

2. **Transducer Power Gain**

$$G_T = \frac{\text{Power delivered to load}}{\text{Power available from source}} = \frac{P_{out}}{P_{avs}} \tag{7.31}$$

   The transducer gain will depend on both the source and load impedances, Z_S and Z_L, as well as the 2-port network parameters. G_T is a measure of the advantage gained by using the 2-port instead of a lossless matching network. Consider Figure 7.8, which shows the system with the original 2-port replaced with a lossless matching network.

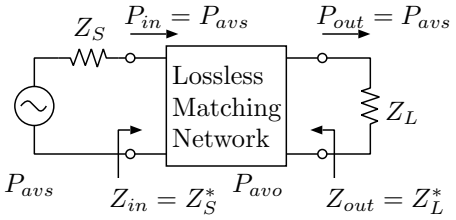


Figure 7.8: Transducer power gain

Since the source "sees" Z_S^*, it will deliver its maximum power, P_{avs}, to the matching network, so $P_{in} = P_{avs}$. Because the matching network is lossless, power delivered to the load will also be P_{avs}, so $P_{out} = P_{avs}$. This means that the transducer gain of a lossless matching network is always equal to 1.

The transducer gain, G_T, calculated for a particular set of source and load terminations measures the performance of the 2-port in that system relative to that of a lossless matching network. If $G_T > 1$, then the power delivered to the load is greater than P_{avs}, which is better than could ever be achieved by using a lossless matching network. On the other hand, if $G_T < 1$, more power would be delivered to the load if the 2-port was replaced with a lossless matching network.

Note that G_T is generally a more useful quantity than is G for system design calculations. To see why this is so, note that it is possible to have a situation where the operating power gain of an amplifier in a particular system is $G = 10$ but the transducer gain in the same system is $G_T = .5$. Here the power gain G looks good since $\frac{P_{out}}{P_{in}} = 10$ – obviously, this 2-port must be contain an "active" device – but G_T says we'd be better off (i.e. more power would be delivered to the load) if the 2-port was replaced with a passive, lossless matching network which, by definition, has $G_T = 1$.

To reconcile the difference between the two gains, take a look at Equation 7.32, which shows that the ratio G_T/G is equal to the fraction of the P_{avs} that is delivered to the input of the 2-port, i.e. the input mismatch factor:

$$\begin{aligned} \frac{G_T}{G} &= \frac{\frac{P_{out}}{P_{avs}}}{\frac{P_{out}}{P_{in}}} = \frac{P_{in}}{P_{avs}} \\ &= \frac{.5}{10} \\ &= 0.05 \end{aligned} \tag{7.32}$$

The problem with this amplifier is the input mismatch — just 5 percent of the available power from the source is actually delivered to the 2-port! The situation could be improved by adding a lossless matching network between the source and the amplifier which would cause all of the available power from the source to be delivered to the input of the amplifier ($P_{in} = P_{avs}$). The addition of the lossless matching network would change the source impedance seen by the amplifier, but this will not change the

operating power gain because G does not depend on Z_S. So G would still be equal to 10 and hence $P_{out} = 10P_{in} = 10P_{avs}$.

The operating power gain, G, only accounts for what happens to the power after it gets in to the 2-port. It doesn't provide any information about what fraction of the available power actually gets in.

3. **Available Power Gain**

$$G_A \equiv \frac{\text{Power available from the output of } 2-\text{port}}{\text{Power available from the source}} = \frac{P_{avo}}{P_{avs}} \tag{7.33}$$

G_A will depend only on the source impedance and the 2-port network parameters. It describes how much power is potentially available from the output of the 2-port.

It may be helpful to summarize how the different power gains are related. First, the transducer gain will always be less than or equal to the operating power gain since

$$\frac{G_T}{G} = \frac{\frac{P_{out}}{P_{avs}}}{\frac{P_{out}}{P_{in}}} = \frac{P_{in}}{P_{avs}} \leq 1 \tag{7.34}$$

Thus, G_T/G is equal to the impedance mismatch factor at the input port, which is equal to one when $P_{in} = P_{avs}$, or when the source is conjugately matched to the input of the 2-port. The ratio given in equation 7.34 can be used to determine how much the power delivered to the 2-port could be increased if a lossless matching network is added between the source and the 2-port.

The transducer gain will always be less than or equal to the available gain since

$$\frac{G_T}{G_A} = \frac{\frac{P_{out}}{P_{avs}}}{\frac{P_{avo}}{P_{avs}}} = \frac{P_{out}}{P_{avo}} \leq 1 \tag{7.35}$$

So G_T/G_A is equal to the impedance mismatch factor at the output port, which is equal to one when the load is conjugately matched to the output of the 2-port. The ratio given in Equation 7.35 can be used to determine how much the power delivered to the load could be increased if a lossless matching network was added between the 2-port and the load.

The operating gain can be greater than or less than the available gain since

$$\frac{G_A}{G} = \frac{\frac{P_{avo}}{P_{avs}}}{\frac{P_{out}}{P_{in}}} = \frac{\frac{P_{in}}{P_{avs}}}{\frac{P_{out}}{P_{avo}}} \tag{7.36}$$

Depending on whether the input mismatch factor (numerator) or the output mismatch factor (denominator) is the smaller, the ratio G_A/G can be either > 1 or < 1. If the input is match is better than the output match, then the ratio in Equation 7.36 will be > 1. On the other hand, if the output match is better than the input match, then the ratio will be < 1.

Finally, it should be noted that only the operating gain and the available gains can be cascaded. That is, when several amplifiers are connected in cascade, the overall operating or available gain of the cascade is simply the product of the operating or available gains of the individual amplifiers. This is not true for the transducer gain.

When cascading gains, it must be kept in mind that the source impedance that terminates the input of the second amplifier is the output impedance of the first amplifier, etc. Consider

the problem of cascading a number of identical amplifiers. Suppose the overall available gain is the quantity of interest. The available gain of the first amplifier depends on the source termination, Z_S. The available gain of the second amplifier depends on the output impedance of the first amplifier which, in turn, depends on the source termination. The available gain of the third amplifier depends on the output impedance of the second amplifier, etc. So the available gain of a cascade of amplifiers depends on the 2-port network parameters of the individual amplifiers as well as the source termination of the first amplifier. Similarly, if the operating power gain of the cascade is desired, it will be found to depend only on the load termination of the last amplifier and the 2-port parameters of all of the individual amplifiers.

7.5 Calculation of Impedance and Gain using the Impedance Matrix

The concepts defined in the previous section will be applied in this section where the input and output impedances and transducer gain are derived for a 2-port represented by its Z-parameters.

Consider a system consisting of a 2-port embedded between a source with impedance Z_S and peak open-circuit voltage V_S and a load with impedance Z_L (Figure 7.9). The power available from the source P_{avs} is defined to be the power that the source would deliver to a conjugately matched load and is

$$P_{avs} = \frac{|V_S|^2}{8R_S} \tag{7.37}$$

The voltage and current at input and output ports are denoted by (V_1, I_1) and (V_2, I_2), respectively, with current chosen to flow into the two port as shown in Figure 7.9.

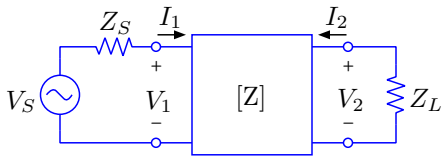


Figure 7.9: A system consisting of source, 2-port, and a load.

The voltage and currents at the input and output of the 2-port are related by impedance matrix elements, Z_{ij}, i.e.

$$V_1 = Z_{11}I_1 + Z_{12}I_2 \tag{7.38}$$

$$V_2 = Z_{21}I_1 + Z_{22}I_2 \tag{7.39}$$

The power delivered to the load impedance is related to P_{avs} by the transducer power gain G_T

$$G_T \equiv \frac{P_{out}}{P_{avs}} = \frac{-\frac{1}{2}\mathrm{Re}[V_2 I_2^*]}{\frac{|V_S|^2}{8R_S}} \tag{7.40}$$

The load constraint $I_2 = -V_2/Z_L$ can be employed in equation 7.40 to write

$$G_T = 4|\frac{V_2}{V_S}|^2\frac{R_S R_L}{|Z_L|^2} \tag{7.41}$$

where $R_S = \text{Re}[Z_S]$ and $R_L = \text{Re}[Z_L]$. Equation 7.41 can be written in terms of the magnitude of the voltage gain, $|V_2/V_1|$, and the voltage division that occurs at the input of the 2-port, $|V_1/V_S|$, i.e.:

$$G_T = 4|\frac{V_2}{V_1}|^2|\frac{V_1}{V_S}|^2\frac{R_S R_L}{|Z_L|^2} \tag{7.42}$$

The voltage division can be written in terms of the input impedance of the 2-port and the source impedance

$$\frac{V_1}{V_S} = \frac{Z_{IN}}{Z_{IN} + Z_S} \tag{7.43}$$

and the input and output impedances of the 2-port can be obtained from the load constraint and equations 7.38 and 7.39. The input impedance is:

$$Z_{IN} = \frac{V_1}{I_1} = Z_{11} - \frac{Z_{12}Z_{21}}{Z_L + Z_{22}} \tag{7.44}$$

and the output impedance is:

$$Z_{OUT} = \frac{V_2}{I_2}|_{V_S=0} = Z_{22} - \frac{Z_{12}Z_{21}}{Z_S + Z_{11}} \tag{7.45}$$

The voltage gain is obtained from 7.38 and 7.39 using the load constraint $I_2 = -V_2/Z_L$:

$$A_V = \frac{V_2}{V_1} = \frac{Z_{21}Z_L}{Z_{11}Z_L + Z_{11}Z_{22} - Z_{12}Z_{21}} \tag{7.46}$$

Now, equation 7.42 can be rewritten using 7.46 and 7.44 as:

$$G_T = 4|\frac{Z_{21}Z_L}{Z_{11}Z_L + Z_{11}Z_{22} - Z_{12}Z_{21}}|^2|\frac{Z_{11} - \frac{Z_{12}Z_{21}}{Z_L+Z_{22}}}{Z_{11} - \frac{Z_{12}Z_{21}}{Z_L+Z_{22}} + Z_S}|^2\frac{R_S R_L}{|Z_L|^2}$$

which simplifies to the final result:

$$G_T = 4\frac{|Z_{21}|^2 R_L R_S}{|(Z_{11} + Z_S)(Z_{22} + Z_L) - Z_{12}Z_{21}|^2} \tag{7.47}$$

Since the available source power P_{avs} is a constant determined by the capabilities of the source, the output power will be proportional to G_T. The largest possible output power with passive source and load terminations results when Z_S and Z_L are chosen to maximize G_T subject to $\text{Re}[Z_S]>0$ and $\text{Re}[Z_L]>0$. With a considerable amount of algebraic manipulation, it can be shown that if a stability factor (usually denoted by "K") is larger than 1, i.e. if

$$K = \frac{2\Re(Z_{11})\Re(Z_{22}) - \Re(Z_{12}Z_{21})}{|Z_{12}Z_{21}|} > 1, \tag{7.48}$$

then a unique set of passive source and load terminations will maximize G_T. The optimum terminations are denoted by Z_{MS} and Z_{ML} and it can also be shown that these terminations

result in a simultaneous conjugate match at the input and output ports, i.e. Z_{MS} and Z_{ML} satisfy the simultaneous conjugate match conditions

$$\begin{aligned} Z_{11} - \frac{Z_{12}Z_{21}}{Z_{ML} + Z_{22}} &= Z_{MS}^* \\ Z_{22} - \frac{Z_{12}Z_{21}}{Z_{MS} + Z_{11}} &= Z_{ML}^* \end{aligned} \quad . \tag{7.49}$$

The significance of the stability factor, K, will be discussed in more detail in Chapter 8.

7.6 Applications of 2-port analysis

7.6.1 Losses in L-networks for impedance matching

Consider an L-network designed to match two resistive terminations, R_S and R_L, with $R_S < R_L$. The design of such networks has been considered previously. Suppose that such a network has been designed and that the series and shunt elements available for use in the network are lossy i.e., suppose that the series and shunt elements have component Q's at the design frequency, ω_o, denoted by Q_s and Q_p, respectively. The loss in the series element can be represented by a series resistance, r_s, and the loss in the parallel element can be represented by a shunt resistance, r_p. The circuit model is shown below in Figure 7.10. The series and shunt reactances are assumed to be those necessary to match the source to the load in the absence of any losses in the network. Thus, the load impedance seen by the resistive 2-port will be equal to $Z_L = R_S - jX_S$.

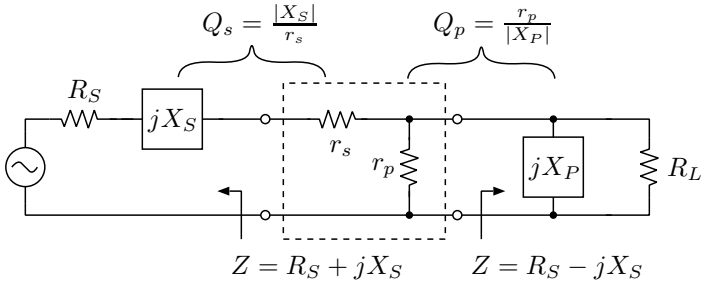


Figure 7.10: Circuit model for an L-network with lossy components. The loss resistances can be lumped into a 2-port which is contained within the dashed box.

As shown in the Figure, the loss resistances can be lumped into a 2-port which is contained within the dashed box. The Z-parameter matrix for the 2-port is easily shown to be:

$$[Z] = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} = \begin{bmatrix} r_s + r_p & r_p \\ r_p & r_p \end{bmatrix} \tag{7.50}$$

Notice that the 2-port defined here is reciprocal and hence $Z_{12} = Z_{21}$. The loss caused by the presence of the resistances r_s and r_p can be determined by calculating the transducer

gain of the 2-port in the system shown in Figure 7.10. The source impedance seen by the resistive 2-port is $Z_S = R_S + jX_S$ and the load impedance is $Z_L = R_S - jX_S$. If the Z-parameters and the terminating impedances are inserted into equation 7.47, the result is:

$$G_T = \frac{4r_p^2 R_S^2}{|(r_s + r_p + R_S + jX_S)(r_p + R_S - jX_S) - r_p^2|^2} \tag{7.51}$$

After some algebraic manipulation, the transducer gain can be rewritten as:

$$G_T = \frac{1}{|1 + \frac{R_S}{2r_p}(1+Q^2) + \frac{r_s}{2R_S} + \frac{r_s}{2r_p}(1-jQ)|^2} \tag{7.52}$$

where the fact that $|X_S| = R_S Q$ has been used to eliminate X_S from the expression. Next, one can make use of the following relationships to write G_T in terms of the L-network Q (Q) and the component Q's of the lossy series and shunt elements:

$$\begin{aligned} \frac{R_S}{2r_p} &= \frac{Q}{2Q_p(1+Q^2)} \\ \frac{r_s}{2R_S} &= \frac{Q}{2Q_s} \\ \frac{r_s}{2r_p} &= \frac{Q^2}{2Q_s Q_p(1+Q^2)} \end{aligned} \tag{7.53}$$

After using 7.53 in 7.52, G_T can be written as:

$$G_T = \frac{1}{|1 + \frac{Q}{2Q_p} + \frac{Q}{2Q_s} + \frac{(1-jQ)Q^2}{2Q_s Q_p(1+Q^2)}|^2} \tag{7.54}$$

This result can be checked by allowing the losses to approach zero ($Q_s \to \infty$, $Q_p \to \infty$) in which case $G_T \to 1$, as expected.

In practical applications, the component Q's (Q_s and Q_p) will be larger than the L-network Q ($Q_s > Q$ and $Q_p > Q$) in which case the last term in the denominator can be neglected compared to the second and third terms. Thus, a good approximation for the transducer gain is:

$$G_T \simeq \frac{1}{(1 + \frac{Q}{2Q_p} + \frac{Q}{2Q_s})^2} \tag{7.55}$$

This result illustrates the fact that component Q's must be much larger than the L-network Q if component losses are to be reasonably small. For example, suppose that it is necessary to keep the matching network loss below 1 dB — this means that $G_T \geq 10^{(-1/10)} = 0.794$. Assuming that the component Q's are equal ($Q_s = Q_p$), then the constraint $G_T \geq 0.794$ leads to $Q_{s,p} \geq 8.2Q$, which means that the L-network must be implemented with components having component Q's at least 8.2 times as large as the L-network Q. Obviously, this will become impractical when the L-network Q is too large. Large L-network Q's are associated with large resistance transformation ratios[1] so that it becomes more difficult to implement an L-network with low losses as the resistance transformation ratio increases.

[1]Recall that the Q of an L-network is $Q = \sqrt{\frac{R_{big}}{R_{small}} - 1}$.

7.6.2 Two-winding Transformers

7.6.2.1 Equivalent Circuit Model for Two-winding Transformers

Transformers are often utilized in both narrow-band and wide-band RF applications. They can be employed for impedance transformation, phase inversion, and dc isolation. Transformers are also utilized as resonant circuit elements.

Consider the two-winding transformer shown in Figure 7.11. Ignoring losses, the equations that describe this device are

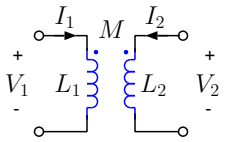


Figure 7.11: 2-winding transformer.

$$V_1 = j\omega L_1 I_1 + j\omega M I_2 \tag{7.56}$$

$$V_2 = j\omega M I_1 + j\omega L_2 I_2 \tag{7.57}$$

where L_1 and L_2 are the self inductances of the transformer windings and M is the mutual inductance. The "dot" convention is such that if current flows into the dotted terminals, the magnetic fluxes linking the two coils will reinforce each other. With this convention, M will be a positive number and will satisfy

$$0 \leq M \leq \sqrt{L_1 L_2}. \tag{7.58}$$

The 2-winding transformer is completely described by its open-circuit impedance matrix,

$$[Z] = \begin{bmatrix} j\omega L_1 & j\omega M \\ j\omega M & j\omega L_2 \end{bmatrix}.$$

Since the circuit operation of the transformer is completely described by this impedance matrix, any network having the same defining equations can be substituted for the transformer. One useful equivalent circuit is shown in Figure 7.12.

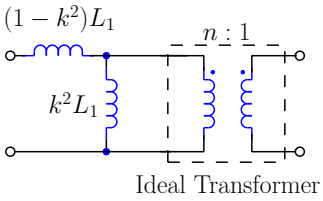


Figure 7.12: Equivalent circuit for a two-winding transformer

In this model the real transformer has been replaced with an ideal transformer and equivalent inductances. The equivalent circuit parameters are k (coupling coefficient), n

(turns ratio), and L_1 (self inductance of winding 1. This equivalent circuit is not unique. For example, it is possible to derive an equivalent circuit that has the self inductance of winding 2 as a parameter. The terminal relations for the ideal transformer are summarized by Figure 7.13 and equations 7.59 and 7.60.

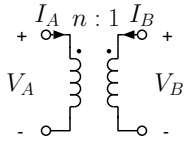


Figure 7.13: Terminal relations for the ideal transformer

$$V_B = \frac{V_A}{n} \tag{7.59}$$

$$I_B = -nI_A \tag{7.60}$$

It is left as an exercise to demonstrate that the equivalent circuit shown in Figure 7.12 has the same impedance matrix as the actual transformer, provided that the following relationships hold:

$$k = \frac{M}{\sqrt{L_1 L_2}} \tag{7.61}$$

$$n = k\sqrt{\frac{L_1}{L_2}} \tag{7.62}$$

The k parameter is called the *coefficient of coupling* ($0 \leq k \leq 1$) and n is the *effective turns ratio* for the transformer.

A useful approximation results when the coefficient of coupling approaches 1 in which case the two windings are said to be *tightly coupled*. This will be the situation when both windings are wound on a high permeability magnetic core. Here the equivalent circuit reduces to the circuit shown in Figure 7.14.

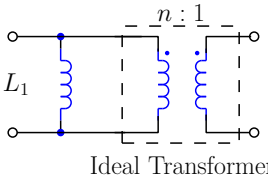


Figure 7.14: Equivalent circuit when k approaches 1. In this case, the effective turns ratio is, approximately, $n = \sqrt{L_1/L_2}$.

7.6.2.2 Impedance Transformation with the Two-winding Transformer

Consider the situation where a tightly coupled transformer is to be used as an impedance transformer as in Figure 7.15.

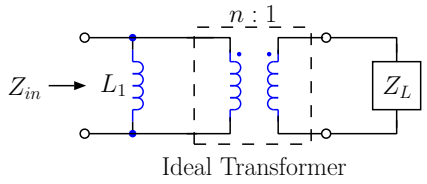


Figure 7.15: Tightly coupled transformer used as impedance transformer

The Z-parameter matrix for a two-winding transformer is

$$[Z] = \begin{bmatrix} j\omega L_1 & j\omega M \\ j\omega M & j\omega L_2 \end{bmatrix} = \begin{bmatrix} j\omega L_1 & j\omega L_1 \frac{k^2}{n} \\ j\omega L_1 \frac{k^2}{n} & j\omega L_1 \frac{k^2}{n^2} \end{bmatrix}.$$

This is the exact impedance matrix. When the transformer is tightly coupled, $k \simeq 1$, and the impedance matrix can be approximated by

$$[Z] \simeq \begin{bmatrix} j\omega L_1 & j\omega L_1/n \\ j\omega L_1/n & j\omega L_1/n^2 \end{bmatrix}.$$

The impedance seen looking in to a tightly coupled transformer that is terminated in load impedance Z_L is now easily computed:

$$Z_{IN} = Z_{11} - \frac{Z_{12}Z_{21}}{Z_{22} + Z_L} = j\omega L_1 - \frac{(j\omega L_1/n)^2}{j\omega L_1/n^2 + Z_L} = \frac{j\omega L_1 Z_L n^2}{j\omega L_1 + Z_L n^2} \tag{7.63}$$

The tightly coupled transformer will behave essentially like an ideal transformer if $\omega L_1 \gg |Z_L n^2|$ or, since $n^2 = L_1/L_2$, if $\omega L_2 \gg |Z_L|$. In this case

$$Z_{IN} \approx n^2 Z_L. \tag{7.64}$$

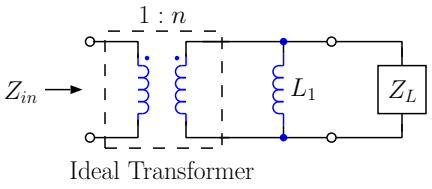


Figure 7.16: Transformer reversed

Similarly, if the transformer is reversed, as in Figure 7.16 then

$$Z_{IN} = \frac{1}{n^2} \frac{s\,L_1\,Z_L}{s\,L_1 + Z_L} \tag{7.65}$$

And if

$$\omega L_1 \gg Z_L \tag{7.66}$$

then

$$Z_{IN} \simeq \frac{1}{n^2} Z_L \tag{7.67}$$

These examples lead to a rule that is applicable when designing tightly coupled impedance transformers: **The inductive reactance of the impedance transformer winding should be significantly larger than the impedance to which it is connected**. Typically, the inductive reactance is chosen to be at least four times larger than the impedance at the lowest frequency of intended operation. When this constraint is satisfied, the tightly coupled transformer can provide an impedance transformation that is essentially independent of frequency over a wide bandwidth.

7.6.2.3 Single-tuned Transformer

In Section 7.6.2.2 the impedance transforming properties of the tightly coupled transformer were examined, and a working rule was given for designing a transformer which will behave essentially like an ideal transformer. This is most useful when designing transformers that must operate over a relatively wide frequency range. If narrow-band operation is desired, then the inductive reactance of the transformer windings can be resonated, and ideal transformer operation will result at the resonant frequency. For example, suppose that a capacitance is used to resonate the primary inductance of the transformer as in Figure 7.17.

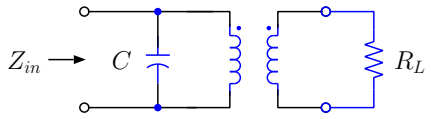


Figure 7.17: Actual circuit

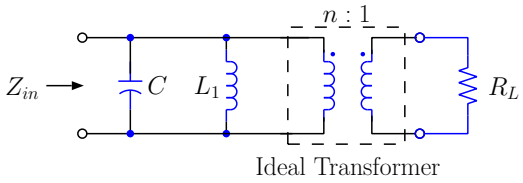


Figure 7.18: Single-tuned transformer using a tightly coupled transformer and a resonating capacitance.

If the capacitor resonates with L_1 at the frequency of interest, then at that frequency, the circuit will behave like an ideal transformer, and the input impedance will be $Z_{in} = n^2 R_L$. More generally, the primary circuit can be modeled by reflecting the load resistance through the transformer as shown in Figure7.19. The result is a parallel resonant circuit with

$$Q = \frac{n^2 R_L}{\omega_o L_1} = \frac{R_L}{\omega_o L_2}$$

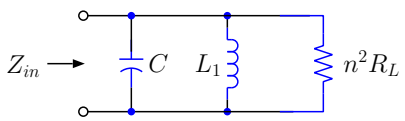


Figure 7.19: Load resistance reflected through transformer

and

$$\omega_o = \frac{1}{\sqrt{L_1 C}}.$$

The input impedance at the resonant frequency is

$$Z_{IN} = n^2 R_L. \tag{7.68}$$

The bandwidth of the single-tuned transformer circuit is the same as the bandwidth of the equivalent parallel RLC circuit shown in Figure 7.19, i.e.,

$$BW = \frac{\omega_o}{Q} = \frac{\omega_o^2 L_1}{n^2 R_L} = \frac{1}{n^2 C R_L}.$$

The bandwidth can be controlled by adjusting the self inductance of the primary winding, L_1, and the turns ratio, n. For a given L_1, the resonating capacitance, C, must be chosen to resonate with L_1 at ω_o. This provides a useful method for designing impedance matching networks with specified bandwidth.

7.6.3 Two Magnetically-Coupled Resonators (Doubly-Tuned Transformer)

The Z matrix for two coupled coils can be written in terms of the self-inductances, L_1 and L_2 and mutual-inductance M:

$$[Z] = \begin{bmatrix} j\omega L_1 & j\omega M \\ j\omega M & j\omega L_2 \end{bmatrix} \tag{7.69}$$

In this section, we'll employ the Z matrix to investigate the properties of a filter consisting of two magnetically coupled resonators. The system is shown in Figure 7.20.

We assume that the two coils are identical ($L_1 = L_2 = L$), and that resistive losses can be modeled by a resistance, r, in series with each of the coils. In addition, a capacitance, C, can be added in series with each coil to form two magnetically-coupled series resonant circuits. It is easy to verify that adding the elements r and C in series with each coil changes the diagonal elements of the Z matrix by adding the impedances of these elements to the self-impedance of the coils and does not affect the off-diagonal (coupling) elements of the Z matrix. Thus, the impedance matrix of the coupled resonator system is given by:

$$[Z] = \begin{bmatrix} r + j\omega L + \frac{1}{j\omega C} & j\omega M \\ j\omega M & r + j\omega L + \frac{1}{j\omega C} \end{bmatrix} \tag{7.70}$$

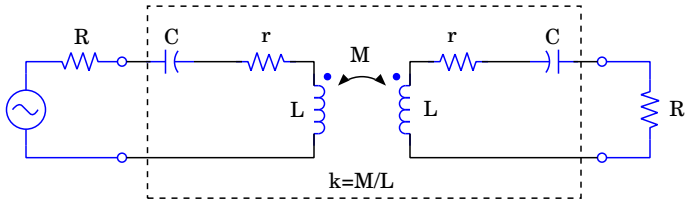


Figure 7.20: Two magnetically-coupled series resonant circuits in a system with source and load impedances equal to R.

In a system with source and load impedance denoted by R, the transducer gain of the coupled-resonator filter is obtained by using 7.70 and $Z_S = R$, $Z_L = R$ in Equation 7.47:

$$G_T = \frac{4\omega^2 M^2 R^2}{|(R + r + j\omega L + \frac{1}{j\omega C})^2 + \omega^2 M^2|^2} \tag{7.71}$$

Define the coupling factor :

$$k = \frac{M}{L}, \tag{7.72}$$

the resonant frequency of either resonator in isolation:

$$\omega_o = \frac{1}{\sqrt{LC}}, \tag{7.73}$$

and the loaded Q of the resonator (accounting for the loss resistance and termination resistance):

$$Q' = \frac{\omega_o L}{R + r}. \tag{7.74}$$

Equations 7.72 - 7.74 can be used in Equation 7.71 to write the transducer gain as:

$$G_T = \frac{R^2}{(R + r)^2} \frac{4k^2 Q'^2 \frac{\omega^2}{\omega_o^2}}{|(1 + jQ'(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}))^2 + k^2 Q'^2 \frac{\omega^2}{\omega_o^2}|^2} \tag{7.75}$$

This is a complicated-looking result, so to simplify interpretation it is useful to examine the transducer gain at the resonant frequency ω_o :

$$G_T|_{\omega=\omega_o} = \frac{R^2}{(R + r)^2} \frac{4k^2 Q'^2}{(1 + k^2 Q'^2)^2} \tag{7.76}$$

In the absence of resonator losses ($r \to 0$) the first term is equal to one. The second term will be equal to one also if

$$k = \frac{1}{Q'} \tag{7.77}$$

Thus (in the absence of resonator losses), if the coupling factor and/or loaded Q are adjusted so that equation 7.77 is satisfied, we expect that the coupled resonator filter will transmit signals unattenuated at the resonant frequency, ω_o. In this case, the resonators are said to be “critically coupled”.

Figure 7.21 shows the transducer gain for a system employing resonators with loaded Q equal to 5. Curves are plotted for 5 different coupling factors, two of which are smaller than the critical coupling value of 1/5=0.2 and two of which are larger than the critical value. Coupled resonators with k smaller than the critical coupling value are said to be "undercoupled", whereas resonators with k larger than the critical value are "overcoupled". A characteristic of the overcoupled response is a double-peaked response with separation between the peaks that depends on the degree of overcoupling. Undercoupling, on the other hand, yields a single-peaked response with progressively larger attenuation as the coupling factor is decreased below the critical value. Notice that the critically coupled response yields a bandpass filter with relatively flat passband response. An obvious disadvantage of this type of filter is the fact that the optimum coupling factor depends on the source and load impedances. Thus, a filter that is operated with different terminations than it was designed for may end up either under- or over-coupled and the shape of the filter response may degrade significantly under such conditions.

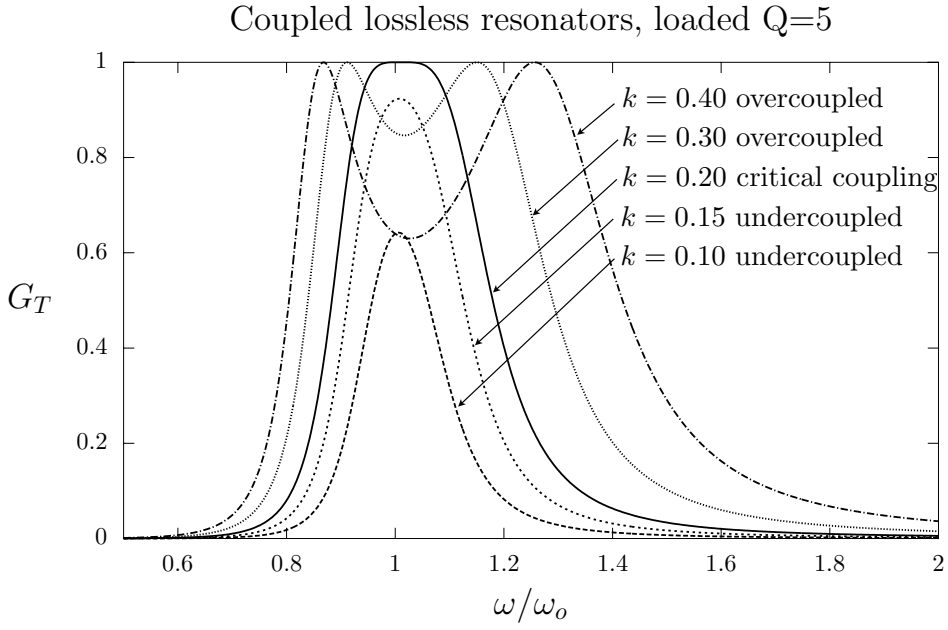


Figure 7.21: Transducer power gain for a coupled resonator filter using resonators with loaded Q equal to 5. Critical coupling will occur when $k = 1/Q' = 1/5 = 0.2$.

7.6.4 Analysis of a Small-signal Series/Shunt Feedback Amplifier

In this section we will calculate the input and output impedance and voltage gain of the amplifier shown in Figure 7.22. The amplifier is redrawn in Figure 7.23 to show how the circuit can be described as a combination of 2-ports connected in series, in cascade, and in parallel. The hybrid-π 2-port is in series with the 2-port containing the single shunt resistor R_e. The output of the resulting 2-port is cascaded with the input of a 2-port consisting of an ideal transformer with 1:1 turns ratio. The ideal transformer is configured so that it produces a 4:1 impedance transformation. The cascade is in parallel with a 2-port consisting of the single series resistor R_f.

Analysis proceeds as follows:

- calculate the Z parameters of the hybrid-π 2-port and the 2-port containing the shunt resistor, R_e. Add these two Z matrices to obtain the Z matrix of the series combination.

- Convert the Z matrix of the series combination to an $ABCD$ matrix. Denote this matrix by $[ABCD]_1$. Determine the $ABCD$ matrix of the $4:1$ transformer and denote this matrix by $[ABCD]_2$. The $ABCD$ matrix of the cascade is then $[ABCD]_1[ABCD]_2$. Note that the order of the terms in this matrix product is important and must reflect the order in which the 2-ports are cascaded.

- Convert $ABCD$ of the cascade to a Y matrix and sum this matrix with the Y matrix for the 2-port consisting of the single series resistor, R_f.. The result is the Y matrix of the overall circuit.

- Calculate Y_{in}, Y_{out}, and A_v in terms of the Y parameters of the overall circuit.

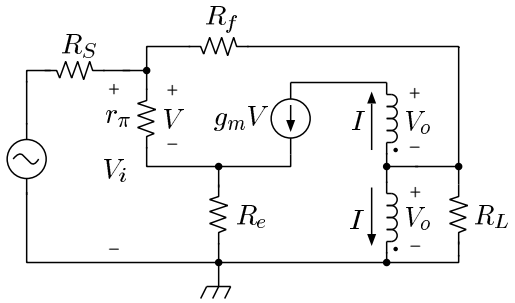


Figure 7.22: Small signal model of amplifier with series and shunt feedback and 4:1 output transformer.

The Z matrix for the hybrid-π 2-port is:

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} = \begin{bmatrix} r_\pi & 0 \\ -g_m r_\pi r_o & r_o \end{bmatrix}$$

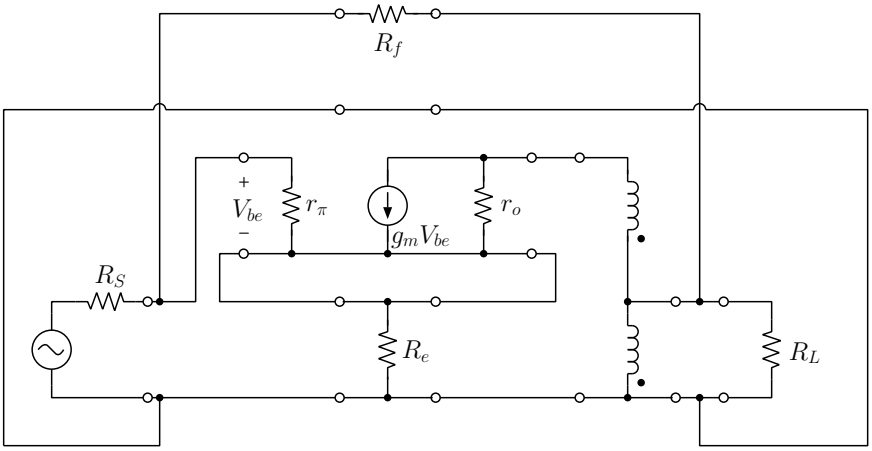


Figure 7.23: Feedback amplifier drawn as a set of interconnected 2-ports.

The Z matrix for the shunt resistor is:

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} = \begin{bmatrix} R_e & R_e \\ R_e & R_e \end{bmatrix}$$

The Z matrix of the series combination of the hybrid-π 2-port and the shunt resistor is:

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} = \begin{bmatrix} r_\pi + R_e & R_e \\ -g_m r_\pi r_o + R_e & r_o + R_e \end{bmatrix}$$

Convert the Z matrix to an $ABCD$ matrix:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \frac{1}{-g_m r_o r_\pi + R_e} \begin{bmatrix} (R_e + r_\pi) & (R_e + r_\pi)(R_e + r_o) - R_e(R_e - g_m r_o r_\pi) \\ 1 & r_o + R_e \end{bmatrix}$$

At this point, we will let $r_o \to \infty$ since the small signal model shown in Figure 7.22 does not contain this parameter. Why didn't we do this in the first place? Because the Z matrix of the hybrid-π model wouldn't have existed! After letting $r_o \to \infty$ the $ABCD$ matrix reduces to:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \frac{-1}{g_m r_\pi} \begin{bmatrix} 0 & R_e + r_\pi + g_m R_e r_\pi \\ 0 & 1 \end{bmatrix}$$

The $ABCD$ matrix of the 4:1 transformer implemented using the ideal transformer with 1:1 turns ratio is:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

The product of the two $ABCD$ matrices is:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \frac{-1}{2g_m r_\pi} \begin{bmatrix} 0 & R_e + r_\pi + g_m R_e r_\pi \\ 0 & 1 \end{bmatrix}$$

Convert the $ABCD$ matrix to a Y matrix:

$$\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} = \frac{2g_m r_\pi}{R_e + r_\pi + g_m R_e r_\pi} \begin{bmatrix} \frac{1}{2g_m r_\pi} & 0 \\ 1 & 0 \end{bmatrix}$$

The Y matrix of the 2-port consisting of the series resistor R_f is:

$$\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} = \frac{1}{R_f} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

The sum of the 2 Y matrices is the Y matrix for the overall circuit:

$$\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{R_e + r_\pi + g_m R_e r_\pi} + \frac{1}{R_f} & -\frac{1}{R_f} \\ \frac{2g_m r_\pi}{R_e + r_\pi + g_m R_e r_\pi} - \frac{1}{R_f} & \frac{1}{R_f} \end{bmatrix}.$$

The input impedance of a 2-port can be written in terms of the Y parameters of the 2-port and the load admittance, Y_L, as:

$$Y_{IN} = Y_{11} - \frac{Y_{12}Y_{21}}{Y_{22} + Y_L}.$$

Insert the Y parameters of the overall 2-port and use $Y_L = \frac{1}{R_L}$:

$$Y_{IN} = \frac{1}{R_e + r_\pi + g_m R_e r_\pi} + \frac{1}{R_f} - \frac{-\frac{1}{R_f}\left(\frac{2g_m r_\pi}{R_e + r_\pi + g_m R_e r_\pi} - \frac{1}{R_f}\right)}{\frac{1}{R_f} + \frac{1}{R_L}}.$$

The input impedance is then found from $Z_{IN} = Y_{IN}^{-1}$. After some simplification, the input impedance can be written as:

$$Z_{IN} = \frac{(R_L + R_f)(R_e + r_\pi + g_m R_e r_\pi)}{R_e + r_\pi + R_L + R_f + g_m r_\pi (R_e + 2R_L)}.$$

Similarly, the output admittance depends on the Y parameters and the source admittance, Y_S:

$$Y_{OUT} = Y_{22} - \frac{Y_{12}Y_{21}}{Y_{11} + Y_S}.$$

Inserting the Y parameters of the overall 2-port, use $Y_S = \frac{1}{R_S}$, and $Z_{OUT} = Y_{OUT}^{-1}$ to show:

$$Z_{OUT} = \frac{(R_S + R_f)(R_e + r_\pi + g_m R_e r_\pi) + R_S R_f}{R_S + R_e + r_\pi + g_m r_\pi (R_e + 2R_S)}.$$

The voltage gain of a 2-port depends on the Y parameters and the load admittance $Y_L = \frac{1}{R_L}$:

$$A_v = \frac{-Y_{21}}{Y_{22} + Y_L}.$$

Insert the Y parameters and simplify to show:

$$A_v = -2g_m \frac{R_L R_f}{R_L + R_f} \frac{1}{1 + R_e(g_m + \frac{1}{r_\pi})} + \frac{R_L}{R_L + R_f}.$$

7.7 Y, Z, h, ABCD relationships

Relationships between the parameter sets described in this Chapter are given in the following sections. The determinants of the parameter matrices are defined as follows:

$$D_Y = Y_{11}Y_{22} - Y_{21}Y_{12} \tag{7.78}$$

$$D_Z = Z_{11}Z_{22} - Z_{21}Z_{12} \tag{7.79}$$

$$D_h = h_{11}h_{22} - h_{21}h_{12} \tag{7.80}$$

$$D_{ABCD} = AD - BC \tag{7.81}$$

7.7.1 Converting to Y-parameters

$$\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} = \begin{bmatrix} \frac{Z_{22}}{D_Z} & -\frac{Z_{12}}{D_Z} \\ -\frac{Z_{21}}{D_Z} & \frac{Z_{11}}{D_Z} \end{bmatrix} = \begin{bmatrix} \frac{1}{h_{11}} & -\frac{h_{12}}{h_{11}} \\ \frac{h_{21}}{h_{11}} & \frac{D_h}{h_{11}} \end{bmatrix} = \begin{bmatrix} \frac{D}{B} & -\frac{D_{ABCD}}{B} \\ -\frac{1}{B} & \frac{A}{B} \end{bmatrix}$$

7.7.2 Converting to Z-parameters

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} = \begin{bmatrix} \frac{Y_{22}}{D_Y} & -\frac{Y_{12}}{D_Y} \\ -\frac{Y_{21}}{D_Y} & \frac{Y_{11}}{D_Y} \end{bmatrix} = \begin{bmatrix} \frac{D_h}{h_{22}} & \frac{h_{12}}{h_{22}} \\ -\frac{h_{21}}{h_{22}} & \frac{1}{h_{22}} \end{bmatrix} = \begin{bmatrix} \frac{A}{C} & \frac{D_{ABCD}}{C} \\ \frac{1}{C} & \frac{D}{C} \end{bmatrix}$$

7.7.3 Converting to h-parameters

$$\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} = \begin{bmatrix} \frac{D_Z}{Z_{22}} & \frac{Z_{12}}{Z_{22}} \\ -\frac{Z_{21}}{Z_{22}} & \frac{1}{Z_{22}} \end{bmatrix} = \begin{bmatrix} \frac{1}{Y_{11}} & -\frac{Y_{12}}{Y_{11}} \\ \frac{Y_{21}}{Y_{11}} & \frac{D_Y}{Y_{11}} \end{bmatrix} = \begin{bmatrix} \frac{B}{D} & \frac{D_{ABCD}}{D} \\ -\frac{1}{D} & \frac{C}{D} \end{bmatrix}$$

7.7.4 Converting to ABCD-parameters

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} \frac{Z_{11}}{Z_{21}} & \frac{D_Z}{Z_{21}} \\ \frac{1}{Z_{21}} & \frac{Z_{22}}{Z_{21}} \end{bmatrix} = \begin{bmatrix} -\frac{Y_{22}}{Y_{21}} & -\frac{1}{Y_{21}} \\ -\frac{D_Y}{Y_{21}} & -\frac{Y_{11}}{Y_{21}} \end{bmatrix} = \begin{bmatrix} -\frac{D_h}{h_{21}} & -\frac{h_{11}}{h_{21}} \\ -\frac{h_{22}}{h_{21}} & -\frac{1}{h_{21}} \end{bmatrix}$$

7.8 Summary

7.8.1 Z parameters

$$Z_{IN} = Z_{11} - \frac{Z_{12}Z_{21}}{Z_L + Z_{22}} \tag{7.82}$$

$$Z_{OUT} = Z_{22} - \frac{Z_{12}Z_{21}}{Z_S + Z_{11}} \tag{7.83}$$

$$A_V = \frac{Z_{21}Z_L}{Z_{11}Z_L + Z_{11}Z_{22} - Z_{12}Z_{21}} \tag{7.84}$$

$$A_I = \frac{-Z_{21}}{Z_{22} + Z_L} \tag{7.85}$$

7.8.2 Y parameters

$$Y_{IN} = Y_{11} - \frac{Y_{12}Y_{21}}{Y_L + Y_{22}} \tag{7.86}$$

$$Y_{OUT} = Y_{22} - \frac{Y_{12}Y_{21}}{Y_S + Y_{11}} \tag{7.87}$$

$$A_V = \frac{-Y_{21}}{Y_{22} + Y_L} \tag{7.88}$$

$$A_I = \frac{Y_{21}Y_L}{Y_{11}Y_L + Y_{11}Y_{22} - Y_{12}Y_{21}} \tag{7.89}$$

7.9 References

1. Terman, Frederick Emmons, *Radio Engineers' Handbook*, McGraw-Hill Book Company, 1943.

2. Krauss, Herbert L., Charles W. Bostian, and Frederick H. Raab, *Solid State Radio Engineering*, John Wiley and Sons, 1980.

3. Ludwig, Reinhold, and Pavel Bretchko, *RF Circuit Design - Theory and Applications*, Prentice-Hall, Inc., New Jersey, 2000.

4. Balabanian, Norman and Theodore A. Bickart, *Electrical Network Theory*, John Wiley & Sons, 1969.

7.10 Homework Problems

1. Suppose the Y-parameters of a 2-port are known.

   (a) Derive an expression for the input admittance of the 2-port when it is terminated with load admittance Y_L.

   (b) Derive an expression for the voltage gain $A_v = V_2/V_1$. Your result will depend on some of the Y-parameters and Y_L.

   (c) Using your results for parts 1a and 1b find an expression for the operating power gain of the 2-port, G. The operating power gain is defined as

   $$G = \frac{P_{out}}{P_{in}} \tag{7.90}$$

   where P_{out} is the power delivered to the load and P_{in} is the power delivered to the 2-port. Your result should be in terms of the Y-parameters of the 2-port and Y_L.

2. Suppose the ABCD-parameters of a 2-port are known.

   (a) Derive an expression for the input impedance of the 2-port when it is terminated with load impedance Z_L.

   (b) Derive an expression for the output impedance of the 2-port when it is terminated with source impedance Z_S.

   (c) Derive an expression for the voltage gain $A_v = V_2/V_1$. Your result will depend on some of the ABCD-parameters and Z_L.

   (d) Find an expression for the transducer power gain, G_T, of the 2-port when it is driven with a source having impedance Z_S and terminated with load impedance Z_L. Your result will depend on the ABCD-parameters as well as Z_S and Z_L. Simplify the expression as much as possible.

3. Find the Z parameters for the 2-port in Figure 7.24.

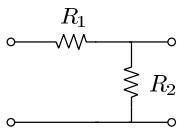


Figure 7.24: 2-port.

4. Find the Z-parameters for the T-network in Figure 7.25. Port 1 is on the left.

5. Find the ABCD parameters for the 2-port shown in Figure 7.24.

6. Consider the unilateral hybrid-pi model shown in Figure 7.26. Find an expression for the available power gain of this 2-port. Your result should be expressed only in terms of Z_1, Z_2, g_m and the impedance of the source Z_s (not shown). Start from the basic definition of available power gain.

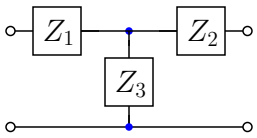


Figure 7.25: T-network.

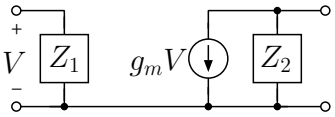


Figure 7.26: Unilateral hybrid-pi model.

7. The small signal equivalent circuit model for an amplifier is shown in Figure 7.27. The amplifier is designed to operate at a center frequency of 30 MHz. Find the 3 dB

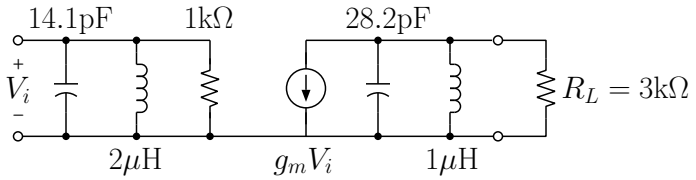


Figure 7.27: Small signal equivalent circuit model for an amplifier.

bandwidth of the **operating power gain**. Express your result in MHz. Start from the basic definition of operating power gain.

8. A lossy network has been designed to provide 10 dB of attenuation (i.e. operating power gain $G = -10\,\text{dB}$) when the network is terminated with a $50\,\Omega$ load. It is also known that the input impedance of the network is $300\,\Omega$ when the output of the network is terminated with a $50\,\Omega$ load. Find the transducer gain for this 2 port in a system with $Z_S = Z_L = 50\,\Omega$.

9. The 2-port shown in Figure 7.28 is characterized by its Z-parameters, where $Z_{11} = Z_{12} = Z_{21} = Z_{22} = R$. Find Z_{in}. Express your result in terms of R only. Hint: The input impedance of a 2-port terminated with load impedance Z_L is $Z_{in} = Z_{11} - \frac{Z_{12}Z_{21}}{Z_{22}+Z_L}$.

10. This exercise will provide some practice in building up a more complex 2-port by combining simple 2-ports in series and in parallel. It is also intended to provide some insight into why we utilize the broadband transformer in the amplifier that is constructed in laboratory 3. Consider the small-signal equivalent circuit shown in the

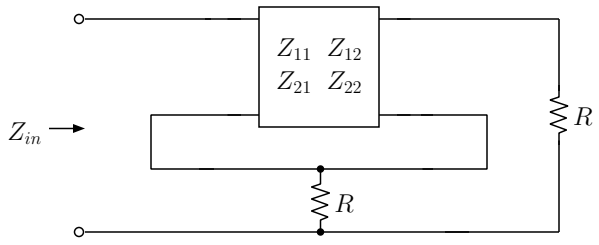
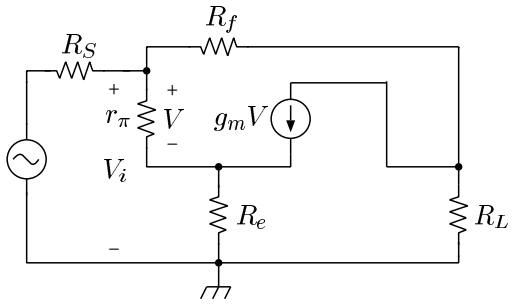


Figure 7.28: Arbitrary 2-port in series with a 2-port consisting of a shunt resistor (a series-feedback configuration).

Figure. The resistances R_S and R_L are the source and load resistances, respectively, and are external to the amplifier. This is a simple transistor amplifier with series (R_e) and shunt (R_f) feedback. It is similar to the amplifier that you will design and build in Lab 3, however it does not include the 1:1 transformer between the load and the transistor's collector.



(a) Find the Y-parameter matrix for the feedback amplifier. Do not include R_S and R_L in the amplifier - they are the external source and load terminations.

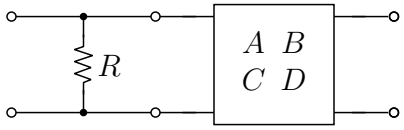
(b) Find expressions for the input impedance (Z_{in}), output impedance (Z_{out}), and voltage gain (A_v) of the amplifier when terminated with R_S and R_L.

(c) Now, follow the analysis given in section 7.6.4 to determine expressions for the feedback resistance R_f, and power gain G under the assumption that the parameters are adjusted to provide a simultaneous (approximate) impedance match at the input and output ports when both ports are terminated with resistance R. Make a table similar to Table 3.1 in the Lab notes. Table 3.1 lists, for each value of R_e, the power gain and the value of R_f required to produce the impedance match assuming that $R_S = R_L = R = 50\,\Omega$, $I_{CQ} = 10\,\text{mA}$, and $\beta = 100$. Table 3.1 was produced using equations 3.12 and 3.14 in the Lab notes. Compare the results for the amplifier analyzed here and the one analyzed in the Lab notes. What advantage (if any) does the addition of the transformer provide?

11. A 2-port is empirically found to have the following properties. (i) With the output port

unterminated the input impedance $Z_{IN} = 100\,\Omega$ and the voltage gain $A_V = \frac{V_2}{V_1} = 0.5$. (ii) With the output port terminated in a short circuit, the input impedance $Z_{IN} = 75\,\Omega$ and the current gain $A_I = \frac{I_2}{I_1} = -0.5$. Find Z_{IN} and A_V when the output port is terminated with $Z_L = 50\,\Omega$.

12. A resistor is added in parallel with the input of an existing 2-port, as shown in the Figure. Determine the overall ABCD matrix of the system. Your result will be expressed



in terms of the ABCD parameters of the original 2-port (denoted by A, B, C, D) and the resistance R.

13. An impedance inverter has the property that the input impedance of the network is equal to the inverse of the load impedance, i.e. $Z_{IN} = Z_L^{-1}$. Determine what properties an ABCD matrix must have to produce the impedance inverter function.

14. Consider the T-network shown in Figure 7.29. The ABCD parameters of the T-

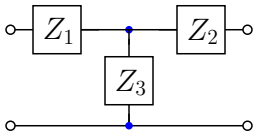


Figure 7.29: T-network to be used as an impedance inverter.

network are:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \frac{1}{Z_3} \begin{bmatrix} Z_1 + Z_3 & Z_1Z_2 + Z_2Z_3 + Z_1Z_3 \\ 1 & Z_2 + Z_3 \end{bmatrix}$$

Determine values of the T-network impedances (Z_1, Z_2, Z_3) that will cause the T-network to be an impedance inverter.

15. A linear, time-invariant, 2-port is empirically determined to have the following properties: (i) with the output port unterminated, the input impedance $Z_{IN} = 500\ \Omega$ and the voltage gain $A_V = 4$. (ii) With the output port terminated in a short circuit, $Z_{IN} = 500\ \Omega$. (iii) With the input port unterminated, the output impedance $Z_{OUT} = 200\ \Omega$. Find the voltage gain, A_V, when the output port is terminated with $Z_L = 50\ \Omega$.

16. An admittance inverter has the property that the input admittance of the 2-port is equal to the inverse of the load admittance, i.e. $Y_{IN} = Y_L^{-1}$.

    (a) Determine the constraints that the Y-parameters of an admittance inverter must satisfy.

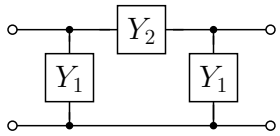


Figure 7.30: Pi-network admittance inverter.

(b) Consider the Pi-network shown in Figure 7.30. Determine the values of the Pi-network admittances (Y_1, Y_2) that will cause the Pi-network to be an admittance inverter.

17. The "h-parameters", h_{ij}, relate the voltages and currents at the input and output of a 2-port according to:

$$V_1 = h_{11}I_1 + h_{12}V_2$$
$$I_2 = h_{21}I_1 + h_{22}V_2$$

(a) Find an expression for the input impedance, Z_{IN}, of a 2-port when the output port is terminated with load impedance Z_L. Express your result in terms of the h-parameters, h_{ij}, and Z_L.

(b) Consider two 2-ports characterized by h-parameter matrices denoted by $\mathbf{H}^1$ and $\mathbf{H}^2$, respectively. Make a sketch that shows how to connect the two 2-ports together to make a new 2-port which will have h-matrix $\mathbf{H} = \mathbf{H}^1 + \mathbf{H}^2$. You may assume that the current flowing into the upper terminal of every port is equal to the current flowing out of the lower terminal of the port.

Chapter 8

2-port Scattering (S) Parameters

8.1 Introduction and Definition of S-parameters

Recall from Chapter 7 that each parameter set discussed therein has a "reference" impedance associated with each port. In the cases discussed so far, the reference impedance is either a short circuit or an open circuit. The four complex numbers that make up the various parameter sets include two parameters that characterize the input and output impedance or admittance of the 2-port when each port is terminated in its reference impedance and two additional parameters that describe the forward and reverse transfer characteristics of the 2-port when terminated with the reference impedance. In each case, the reference impedance is simply the termination that forces one of the independent parameters (e.g. voltage or current) to zero. For example, the reference impedance for the Y-parameters is a short circuit because terminating a port with a short circuit forces the voltage at that port to zero.

At high frequencies it is difficult to implement terminations that accurately represent short- or open-circuits over a broad frequency range. Thus, it is desirable to define a parameter set that is based on a finite reference impedance that is easily realized in practice. In most applications, a reference impedance that is purely resistive and having a moderate value is relatively easy to implement. In order to retain the feature that terminating a port with the reference impedance will force the independent variable corresponding to that port to zero, any parameter set the employs a finite reference impedance must be based on independent variables that are linear combinations of voltage and current rather than just voltage or current. For example, consider the fact that terminating the output port of a 2-port with some resistance R forces $V_2/I_2 = -R$, by Ohm's Law (the minus sign arises because positive current is defined to flow in to the port). Thus, when the output port is terminated with R, we have $V_2 + RI_2 = 0$. Any constant times the quantity $(V_2 + RI_2)$ could be chosen as the independent variable in a parameter set based on a reference impedance R.

Scattering parameters (or S-parameters) can be viewed as a parameter set that is based on a finite reference impedance. In most cases the reference impedance is either $50\,\Omega$ or $75\,\Omega$ and is the same for both ports, although in some cases it is advantageous to allow the reference impedance to be different at port 1 and port 2. In addition to being easier to implement than short- or open-circuits, at HF, VHF, and microwave frequencies 2-ports are much less likely to oscillate (become unstable) when terminated in a finite resistance or

conductance.

Define $Z_o\,(= R_o)$ to be the reference termination for the S-parameter set. Throughout our discussion we will assume that the reference termination is purely resistive and that the same reference termination is used at both ports. Both of these assumptions can be relaxed in a more advanced treatment of scattering parameters. Figure 8.1 shows the voltage and current conventions that we will use for the S-parameter definitions:

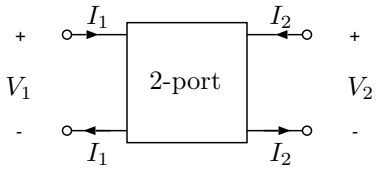


Figure 8.1: Voltage and current conventions for S-parameters.

The S-parameter set is defined in terms of variables (a_1, b_1) (a_2, b_2) which are simply linear combinations of the voltage and current at a particular port, i.e.,

$$a_1 = \frac{V_1 + Z_o I_1}{2\sqrt{Z_o}} \tag{8.1}$$

$$a_2 = \frac{V_2 + Z_o I_2}{2\sqrt{Z_o}} \tag{8.2}$$

$$b_1 = \frac{V_1 - Z_o I_1}{2\sqrt{Z_o}} \tag{8.3}$$

$$b_2 = \frac{V_2 - Z_o I_2}{2\sqrt{Z_o}} \tag{8.4}$$

$$a_1 + b_1 = V_1/\sqrt{Z_o} \tag{8.5}$$

$$a_2 + b_2 = V_2/\sqrt{Z_o} \tag{8.6}$$

$$a_1 - b_1 = \sqrt{Z_o}\, I_1 \tag{8.7}$$

$$a_2 - b_2 = \sqrt{Z_o}\, I_2 \tag{8.8}$$

The independent variables are a_1 and a_2 and the dependent variables are b_1 and b_2. Remember, the reference impedance for a parameter set is that impedance which will force one of the independent variables to zero when it is used to terminate the 2-port. The choice of a finite reference impedance, Z_o, forces us to choose the independent variables to be proportional to $V_i + Z_o I_i$, as defined in Equations 8.1 and 8.2, so that they will be forced to zero when the 2-port is terminated in the reference impedance, Z_o.

The variables defined in Equations 8.1-8.4 have been normalized with respect to the square root of the reference impedance, Z_o. With this normalization (and assuming that the voltages and currents are rms values) the time-averaged power delivered to a given port (denoted by i) can be written as $P = Re[V_i I_i^*] = |a_i|^2 - |b_i|^2$. Thus, the variables have units of [Watts]$^{1/2}$.

Given the definitions in Equations 8.1-8.8, the scattering parameters are defined as follows:

$$b_1 = S_{11}\, a_1 + S_{12}\, a_2 \tag{8.9}$$

$$b_2 = S_{21}\, a_1 + S_{22}\, a_2 \tag{8.10}$$

Thus

$$S_{11} = \frac{b_1}{a_1}|_{a_2=0} \tag{8.11}$$

$$S_{12} = \frac{b_1}{a_2}|_{a_1=0} \tag{8.12}$$

$$S_{21} = \frac{b_2}{a_1}|_{a_2=0} \tag{8.13}$$

$$S_{22} = \frac{b_2}{a_2}|_{a_1=0} \tag{8.14}$$

8.2 Interpretation of S-parameters

Consider S_{11}:

$$S_{11} = \frac{b_1}{a_1}|_{a_2=0} \tag{8.15}$$

Setting $a_2 = 0$ is equivalent to setting

$$V_2 = -Z_o\, I_2$$

$$\frac{V_2}{I_2} = -Z_o \tag{8.16}$$

That is, to set $a_2 = 0$, we terminate port 2 with the reference impedance, Z_o. Thus

$$S_{11} = \frac{b_1}{a_1}|_{a_2=0} = \frac{V_1 - Z_o I_1}{V_1 + Z_o I_1}|_{\text{output port terminated in } Z_o} \tag{8.17}$$

$$= \frac{V_1/I_1 - Z_o}{V_1/I_1 + Z_o}|_{\text{output port terminated in } Z_o}$$

or,

$$S_{11} = \frac{Z_{in} - Z_o}{Z_{in} + Z_o}|_{\text{output port terminated in } Z_o} \tag{8.18}$$

where Z_{in} is the input impedance of the 2-port when the output is terminated in Z_o.

Recall from transmission line theory that S_{11} has the same form as the reflection coefficient, Γ, that relates the incident and reflected voltage waves on a transmission line

$$\Gamma = \frac{V^-}{V^+} = \frac{\text{reflected voltage wave}}{\text{incident voltage wave}} \tag{8.19}$$

where the total voltage on the line, V, is given by

$$V = V^- + V^+ \tag{8.20}$$

Comparison of Equations 8.19 and 8.20 with

$$S_{11} = \frac{b_1}{a_1} \tag{8.21}$$

and

$$V_1 = (a_1 + b_1)\sqrt{Z_o} \tag{8.22}$$

leads to the following correspondence between the variables a_1 and b_1 and the voltage waves on a transmission line:

$$a_i = \frac{1}{\sqrt{Z_o}}(\text{Voltage wave incident on port i}) \tag{8.23}$$

$$b_i = \frac{1}{\sqrt{Z_o}}(\text{Voltage wave reflected from port i}) \tag{8.24}$$

This is an interesting observation, but where is the transmission line? Although there has been no mention of transmission lines in the system so far, it is helpful to imagine that the source and load are connected to the 2-port through sections of transmission line with characteristic impedance Z_o. We'll assume that the sections of transmission line have infinitesimal length, so that they do not affect the electrical characteristics of the system, as in Figure 8.2.

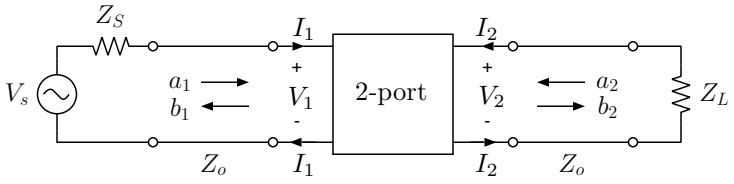


Figure 8.2: Source and load connected to 2-port through transmission line with impedance Z_o

Employing this conceptual model, the a and b variables may be interpreted as representing the (normalized) voltage waves that would exist on the sections of transmission line. When working with S parameters, the model can be used to visualize the a and b variables as incident and reflected (normalized) voltage waves, although in practice the terminations may be connected directly to the 2-port without any intervening transmission lines. (Actually, in practice you will find that it is virtually impossible to get signals into and out of a

2-port without employing transmission lines between the source and the input port and the output port and the load.)

The model shown in Figure 8.2 makes it clear that setting $Z_L = Z_o$ will make $a_2 = 0$, since there will be no reflection from the load termination. Notice that setting $Z_s = Z_o$ would not make $a_1 = 0$ in the circuit shown in Figure 8.2, because the source would cause a non-zero incident wave (a_1) to exist on the input line.

Now consider S_{21}:

$$S_{21} = \frac{b_2}{a_1}|_{a_2=0} \tag{8.25}$$

As discussed earlier, to set $a_2 = 0$, terminate port 2 with Z_o. To aid in obtaining an intuitive feel for the significance of S_{21}, it is helpful to consider the special circuit shown in Figure 8.3 where the source impedance is assumed to be equal to the reference impedance.

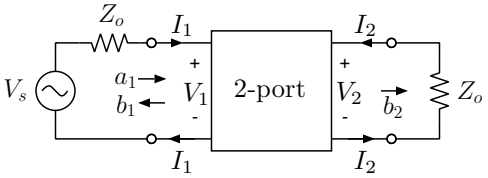


Figure 8.3: A 2-port embedded in a system with $Z_S = Z_L = Z_o$.

Note that the circuit has been drawn without showing any intervening lengths of transmission line between the 2-port and the terminations. The a's and b's can be thought of as existing at the node where the the 2-port is connected to the terminations. To solve for S_{21}, first relate b_2 to the output voltage using Equation 8.4 and the auxiliary relation $V_2/I_2 = -Z_o$:

$$\begin{aligned} b_2 &= \frac{1}{2\sqrt{Z_o}}\left(V_2 - Z_o\, I_2\right) \\ &= V_2/\sqrt{Z_o} \end{aligned} \tag{8.26}$$

Then a_1 can be related to the open circuit voltage of the source

$$a_1 = \frac{1}{2\sqrt{Z_o}}\left(V_1 + Z_o\, I_1\right) \tag{8.27}$$

using

$$I_1 = (V_s - V_1)/Z_o \tag{8.28}$$

Thus

$$\begin{aligned} a_1 &= \frac{1}{2\sqrt{Z_o}}\left(V_1 + V_s - V_1\right) \\ &= V_s/2\sqrt{Z_o} \end{aligned} \tag{8.29}$$

Finally,

$$\begin{aligned} S_{21} &= \frac{b_2}{a_1}|_{a_2=0} = \frac{V_2}{V_s/2}|_{Z_L=Z_o} \end{aligned} \tag{8.30}$$

So for the circuit shown in Figure 8.3, S_{21} is the voltage across the load divided by 1/2 of the open circuit source voltage. Notice that if the 2-port is removed and the load connected directly to the source, then the voltage across the load would be $V_s/2$. In other words, S_{21} is the "insertion voltage gain" in a system where the source and load impedances are both Z_o. Remember that this result is not the definition of S_{21}; it is a special result that was derived *from* the definition for the case when both the source and load impedances are equal to Z_o. This result is often useful when it is necessary to derive an expression for S_{21} starting from the circuit model of a 2-port. S_{12} has a similar interpretation - it is the reverse insertion voltage gain, i.e. the insertion gain that would be measured if the two port was inserted into the system backwards, i.e. with the output port connected to the source and input port connected to the load. The insertion gain interpretion is useful to remember when trying to get an intuitive feel for published or measured values of S_{21} or S_{12}.

8.2.1 Example - Computing S-parameters for a given circuit model

The 2-port in Figure 8.4 consists of a series impedance. To find S_{11}, terminate the 2-port

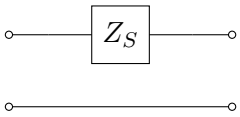


Figure 8.4: 2-port consisting of a series impedance.

in Z_o and find Z_{in} as in Figure 8.5.

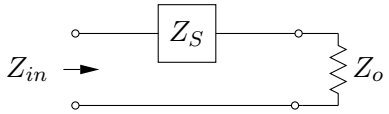


Figure 8.5: Setup for finding S_{11}.

$$Z_{in} = Z_s + Z_o \tag{8.31}$$

$$S_{11} = \frac{Z_{in} - Z_o}{Z_{in} + Z_o} = \frac{Z_s}{Z_s + 2Z_o} \tag{8.32}$$

Because the circuit is not changed if the two ports are reversed, $S_{22} = S_{11}$.

To find S_{21}, refer to Figure 8.6.

$$V_2 = V_s \frac{Z_o}{2Z_o + Z_s} \tag{8.33}$$

$$\frac{V_2}{V_s} = \frac{Z_o}{2Z_o + Z_s}$$

$$S_{21} = \frac{V_2}{V_s/2} = \frac{2Z_o}{2Z_o + Z_s} \tag{8.34}$$

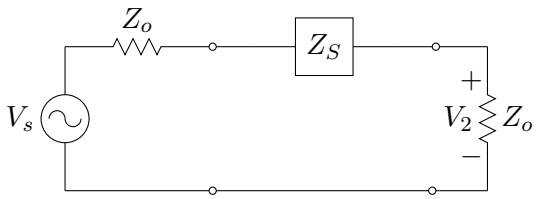


Figure 8.6: Setup for finding S_{21}. Embed the 2-port in a Z_o system, and then $S_{21} = 2V_2/V_s$.

Again, because the circuit is not changed if port 1 and port 2 are reversed, $S_{12} = S_{21}$.

The procedure illustrated in the preceeding example can be applied to any 2-port, and provides a general method for calculating the S-parameters when a circuit model is available.

8.2.2 Summary of 2-port S-parameters:

S_{11} = $\frac{Z_{in}-Z_o}{Z_{in}+Z_o}|_{Z_L=Z_o}$ Input reflection coefficient with output terminated in Z_o

S_{22} = $\frac{Z_{out}-Z_o}{Z_{out}+Z_o}|_{Z_S=Z_o}$ Output reflection coefficient with input terminated in Z_o

S_{21} = Forward insertion voltage gain when source and load impedances are Z_o

S_{12} = Reverse insertion voltage gain when source and load impedances are Z_o

8.2.3 Special relationships for reciprocal, and lossless 2-ports

8.2.3.1 Reciprocal 2-ports

As discussed in Chapter 7, any 2-port that contains no sources, and is constructed only of resistors, capacitors, inductors, transformers, and transmission lines will satisfy reciprocity. The $[S]$ matrix of a reciprocal 2-port will be symmetric, i.e. $S_{12} = S_{21}$.

8.2.3.2 Lossless 2-ports

A lossless 2-port contains no dissipative elements, and hence the time-averaged real power absorbed by the network is zero. A 2-port network constructed only of lossless transmission lines, transformers, and lossless L's and C's will be reciprocal and lossless. As noted in Chapter 7, the constraint that the power absorbed by the network is zero can be written as

$$\Re(V_1 I_1^*) + \Re(V_2 I_2^*) = 0. \tag{8.35}$$

In terms of the variables used to define the scattering parameters, this constraint becomes:

$$|a_1|^2 - |b_1|^2 + |a_2|^2 - |b_2|^2 = 0. \tag{8.36}$$

Use the definitions of the S-parameters to write the b_i in terms of the a_i, then equation 8.36 can be written as

$$|a_1|^2 + |a_2|^2 = (|S_{11}|^2 + |S_{21}|^2)|a_1|^2 + (|S_{12}|^2 + |S_{22}|^2)|a_2|^2 + 2\Re\{(S_{11}S_{12}^* + S_{21}S_{22}^*)a_1 a_2^*\}.$$

Since this constraint must be true for any possible excitation, it must hold for $a_1 = 0$ and for $a_2 = 0$. Setting $a_1 = 0$ we find that

$$|S_{12}|^2 + |S_{22}|^2 = 1. \tag{8.37}$$

Setting $a_2 = 0$ we find that

$$|S_{11}|^2 + |S_{21}|^2 = 1. \tag{8.38}$$

Hence, for arbitrary a_1 and a_2 we must also have

$$S_{11}S_{12}^* + S_{21}S_{22}^* = 0. \tag{8.39}$$

These three relationships say that the dot product of a column of $[S]$ with the conjugate of the same column is unity, whereas the dot product of a column of $[S]$ with the conjugate of the other column is zero. Another way to write these relationships is

$$[S][S]^\dagger = [I],$$

where $[\,]^\dagger$ denotes the conjugate transpose and $[I]$ is the identity matrix.

8.2.3.3 Reciprocal and lossless 2-ports

If a 2-port is reciprocal and lossless, then $S_{12} = S_{21}$ and the dot products of the rows of $[S]$ will satisfy the same constraints as do the columns, in which case the $[S]$ matrix is unitary, i.e.

$$[S][S]^\dagger = [S]^\dagger[S] = [I].$$

8.3 Applications of Scattering Parameters

Consider in Figure 8.7 a 2-port with arbitrary source and load terminations (Z_S, Z_L).

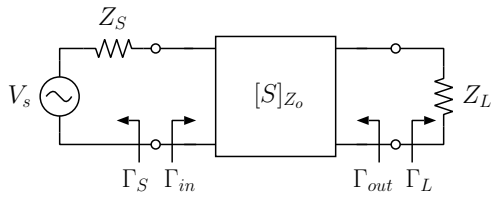


Figure 8.7: 2-port with arbitrary source and load terminations

The reference impedance is assumed to be Z_o. Define Γ_{in} and Γ_{out} to be the input and output reflection coefficients, i.e.,

$$\Gamma_{in} = \frac{Z_{in} - Z_o}{Z_{in} + Z_o} \tag{8.40}$$

$$\Gamma_{out} = \frac{Z_{out} - Z_o}{Z_{out} + Z_o} \tag{8.41}$$

The source reflection coefficient, Γ_S, is defined as

$$\Gamma_S = \frac{Z_S - Z_o}{Z_S + Z_o} \tag{8.42}$$

and the load reflection coefficient, Γ_L, is defined as

$$\Gamma_L = \frac{Z_L - Z_o}{Z_L + Z_o} \tag{8.43}$$

8.3.1 Derivation of Input and Output Reflection Coefficients and Voltage and Current Gains

Consider the input reflection coefficient, $\Gamma_{in} = b_1/a_1$. We expect Γ_{in} to be a function of the four S-parameters as well as the load reflection coefficient, Γ_L. We start by noting that the load termination imposes the following constraint:

$$\frac{a_2}{b_2} = \Gamma_L \tag{8.44}$$

This form of the load constraint may be derived by starting with the voltage/current form:

$$Z_L = -\frac{V_2}{I_2} = -Z_o \frac{a_2 + b_2}{a_2 - b_2} = -Z_o \frac{\frac{a_2}{b_2} + 1}{\frac{a_2}{b_2} - 1}. \tag{8.45}$$

Solve for $\frac{a_2}{b_2}$:

$$\frac{a_2}{b_2} = \frac{Z_L - Z_o}{Z_L + Z_o} = \Gamma_L. \tag{8.46}$$

Use Equation 8.44 and Equation 8.10 to write

$$a_2 = a_1 \frac{S_{21}\,\Gamma_L}{1 - S_{22}\,\Gamma_L} \tag{8.47}$$

Inserting Equation 8.47 into Equation 8.9 yields

$$\Gamma_{in} = \frac{b_1}{a_1} = S_{11} + \frac{S_{12}\,S_{21}\Gamma_L}{1 - S_{22}\,\Gamma_L} \tag{8.48}$$

Similarly, it can be shown that the output reflection coefficient is given by

$$\Gamma_{out} = \frac{b_2}{a_2} = S_{22} + \frac{S_{12}\,S_{21}\,\Gamma_S}{1 - S_{11}\,\Gamma_S} \tag{8.49}$$

Inspection of Equation 8.48 shows that the input reflection coefficient reduces to S_{11} if the load reflection coefficient is 0 ($\Gamma_L = 0$). This is consistent with the discussion in the previous section regarding the interpretation of S_{11}. Similarly, $\Gamma_{out} = S_{22}$ if $\Gamma_S = 0$. In general, however, the input and output reflection coefficients depend on the way the 2-port is terminated. In the special case of a 2-port with $S_{12} = 0$, the input and output reflection coefficients are independent of the terminations. Such a 2-port is said to be unilateral, since the device does not exhibit any reverse transmission. In other words, if a unilateral 2-port ($S_{12} = 0$) is excited at port 2, no response will result at port 1. Although it is usually not possible to construct a 2-port that is perfectly unilateral, in some cases the S_{12} coefficient

is small enough that the 2-port can be considered to be approximately unilateral. The case where $S_{21} = 0$ is also unilateral but is usually not useful, since this 2-port would not give a response at the output when excited at the input.

The voltage gain is:

$$A_V = \frac{V_2}{V_1} = \frac{a_2 + b_2}{a_1 + b_1} = \frac{\frac{a_2}{b_2} + 1}{\frac{a_1}{b_2} + \frac{b_1}{b_2}}. \tag{8.50}$$

Use Equation 8.44 in Equation 8.10:

$$\frac{a_1}{b_2} = \frac{1}{S_{21}}(1 - \Gamma_L S_{22}). \tag{8.51}$$

Then use 8.44 and 8.51 in Equation 8.9 to show that

$$\frac{b_1}{b_2} = \frac{S_{11}}{S_{21}}(1 - \Gamma_L S_{22}) + S_{12}\Gamma_L. \tag{8.52}$$

Finally, use 8.44, 8.51, and 8.52 in 8.50 to obtain the final result:

$$A_V = \frac{S_{21}(1 + \Gamma_L)}{(1 + S_{11})(1 - \Gamma_L S_{22}) + S_{12}S_{21}\Gamma_L}. \tag{8.53}$$

Similarly, the current gain is:

$$A_I = \frac{I_2}{I_1} = \frac{a_2 - b_2}{a_1 - b_1} = \frac{\frac{a_2}{b_2} - 1}{\frac{a_1}{b_2} - \frac{b_1}{b_2}}. \tag{8.54}$$

Use 8.44, 8.51, and 8.52 in 8.54 to obtain the final result:

$$A_I = \frac{S_{21}(\Gamma_L - 1)}{(1 - S_{11})(1 - \Gamma_L S_{22}) - S_{12}S_{21}\Gamma_L}. \tag{8.55}$$

8.3.2 Stability of 2-ports

Before detailed design calculations based on a particular 2-port are made, it is usually necessary to investigate whether the 2-port is potentially unstable. Of course, stability is only a concern when the 2-port contains active elements such as a transistor or a negative resistance device. The questions that need to be answered are:

1. Is the 2-port unconditionally stable? That is, is there any combination of passive source and load terminations (Γ_L, Γ_S) for which the 2-port will oscillate? If not, the 2-port is said to be unconditionally stable.

2. If the 2-port is not unconditionally stable, i.e., if it is potentially unstable, then we would like to be able to determine which source and load terminations make it unstable. If the 2-port is to be used as an amplifier, we would avoid the unstable terminations. If the goal is to design an oscillator, the designer would deliberately choose source and load terminations to cause oscillation.

The stability question can be studied using the negative resistance concept. Considering Figure 8.8, suppose that one of the impedances represents the input or output impedance of the 2-port.

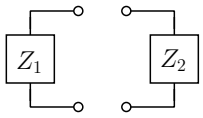


Figure 8.8: Z_1 or Z_2 represents input or output impedance of 2-port

Remember that the circuit will oscillate when the two impedances are connected if Re $[Z_1 + Z_2] \leq 0$ and Im $[Z_1 + Z_2] = 0$. Thus, to make an oscillator, one of the Z_i's must have a negative real part.

For a particular 2-port, if there is some load impedance that makes the real part of $Z_{in} \leq 0$, or if there is some source impedance that makes the real part of $Z_{out} \leq 0$, then the 2-port is potentially unstable because it is possible to choose a passive source or load termination that will make the system oscillate.

An impedance with a negative resistive part corresponds to a reflection coefficient with a magnitude greater than 1. For example, consider Z_{in} and write $Z_{in} = R_{in} + j\,X_{in}$, then

$$\Gamma_{in} = \frac{Z_{in} - Z_o}{Z_{in} + Z_o} \qquad (Z_o = R_o \text{ (real)}) \tag{8.56}$$

$$= \frac{(R_{in} - R_o) + jX_{in}}{(R_{in} + R_o) + jX_{in}}$$

$$|\Gamma_{in}| = \left[\frac{(R_{in} - R_o)^2 + X_{in}^2}{(R_{in} + R_o)^2 + X_{in}^2}\right]^{1/2} \tag{8.57}$$

Examination of Equation 8.57 leads to the following observations:

$$\text{If } R_{in} > 0 \quad \Rightarrow \quad |\Gamma_{in}| < 1 \tag{8.58}$$

$$\text{If } R_{in} \leq 0 \quad \Rightarrow \quad |\Gamma_{in}| \geq 1 \tag{8.59}$$

To determine whether a 2-port is potentially unstable, we check to see if either $|\Gamma_{in}|$ or $|\Gamma_{out}|$ can be larger than or equal to unity.

First consider $|\Gamma_{in}|$:

$$|\Gamma_{in}| = |S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L}| \tag{8.60}$$

$$= |\frac{S_{11} - \Gamma_L(S_{11}S_{22} - S_{12}S_{21})}{1 - S_{22}\Gamma_L}|$$

Define the determinant of the S-parameter matrix, D, as follows:

$$D = S_{11}S_{22} - S_{12}S_{21} \tag{8.61}$$

Then Equation 8.60 becomes

$$|\Gamma_{in}| = |\frac{S_{11} - \Gamma_L D}{1 - S_{22}\Gamma_L}| \tag{8.62}$$

We can now set $|\Gamma_{in}| = 1$ and solve for the corresponding locus of points in the Γ_L plane, i.e., we can solve for the values of Γ_L that make $|\Gamma_{in}| = 1$. Setting $|\Gamma_{in}| = 1$:

$$|S_{11} - \Gamma_L D| = |1 - S_{22}\Gamma_L| \tag{8.63}$$

The absolute value signs can be eliminated by squaring both sides of Equation 8.63:

$$(S_{11} - \Gamma_L D)(S_{11}^* - \Gamma_L^* D^*) = (1 - S_{22}\Gamma_L)(1 - S_{22}^*\Gamma_L^*) \tag{8.64}$$

Expanding both sides of the equation and collecting terms yields

$$|\Gamma_L|^2 + \frac{2\,Re(\Gamma_L(S_{22} - DS_{11}^*))}{|D|^2 - |S_{22}|^2} + \frac{|S_{11}|^2 - 1}{|D|^2 - |S_{22}|^2} \tag{8.65}$$

In deriving Equation 8.65, use has been made of the following identity

$$z + z^* = 2\mathrm{Re}(z) \tag{8.66}$$

where the operator Re() extracts the real part of its argument. Also note that

$$z - z^* = 2\mathrm{Im}(z) \tag{8.67}$$

where the operator Im() extracts the imaginary part of its argument. Now, it is convenient to rewrite Equation 8.65 in terms of the real and imaginary parts of Γ_L.

Let $\Gamma_L = U_L + jV_L$ and substitute Equation 8.65. After some fairly extensive algebraic manipulation, Equation 8.65 can be written in the form

$$(U_L - U_{CL})^2 + (V_L - V_{CL})^2 = {r_L}^2 \tag{8.68}$$

where

$$U_{CL} = \frac{Re(DS_{11}^* - S_{22})}{|D|^2 - |S_{22}|^2} \tag{8.69}$$

$$V_{CL} = \frac{Im(S_{22} - DS_{11}^*)}{|D|^2 - |S_{22}|^2} \tag{8.70}$$

$$r_L = \frac{|S_{12}S_{21}|}{||S_{22}|^2 - |D|^2|} \tag{8.71}$$

Equation 8.68 indicates that the locus of points in the Γ_L plane that correspond to input reflection coefficients with unit magnitude is a circle with center at the complex point $C_L = U_{CL} + jV_{CL}$ and radius r_L. It is convenient to write the coordinates of the center of the stability circle as a complex number, i.e. define $C_L = U_{CL} + jV_{CL}$. Then

$$C_L = \frac{S_{22}^* - D^* S_{11}}{|S_{22}|^2 - |D|^2} \tag{8.72}$$

The circle, when plotted on the Γ_L plane, is referred to as the Γ_L-plane stability circle. Consider the stability circle shown in Figure 8.9.

Figure 8.9 contains two circles. The circle that is centered on the origin is a unit circle (circle with radius = 1) and represents the outer boundary of the Smith Chart. Points within that circle correspond to load reflection coefficients (Γ_L) with magnitudes less than 1 or, equivalently, load impedances with a positive real part. The other circle is the stability circle, and values of Γ_L which lie on that circle will map to input reflection coefficients with magnitude equal to 1 or, equivalently, input impedances that are purely reactive. The stability circle represents the boundary between the region of the Γ_L plane that maps to

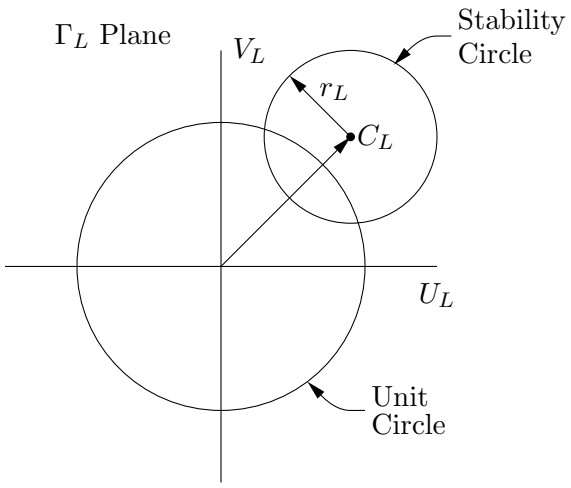


Figure 8.9: Γ_L-plane stability circle

input reflection coefficients with magnitude less than one ($|\Gamma_{in}| < 1$) and the region that maps to ($|\Gamma_{in}| > 1$). The region of the Γ_L plane that maps to ($|\Gamma_{in}| > 1$) is called the unstable region of the Γ_L plane. The two regions (stable and unstable) correspond to the regions inside and outside the stability circle. To decide whether the unstable region corresponds to the region inside or outside of the stability circle, it is sufficient to examine the value of S_{11} for the 2-port under consideration. Note carefully that the origin of the Γ_L plane (the point $\Gamma_L = 0$) maps to S_{11}, i.e., recall

$$\Gamma_{in} = S_{11} + \frac{S_{12}\, S_{21}\, \Gamma_L}{1 - S_{22}\Gamma_L} \tag{8.73}$$

so that when $\Gamma_L = 0$, we have

$$\Gamma_{in} = S_{11} \tag{8.74}$$

Now consider Figure 8.9 and suppose that it is known that $|S_{11}| < 1$. The origin in the Γ_L plane is outside the stability circle and this point maps to S_{11}. This leads us to conclude: (i) the region of the Γ_L plane that lies outside the stability circle maps to $|\Gamma_{in}| < 1$ and (ii) the region inside the stability circle maps to $|\Gamma_{in}| > 1$. The region inside the stability circle would therefore be referred to as the unstable region of the Γ_L plane.

Thus far we have treated only the so-called Γ_L-plane stability circle. A complete characterization of the 2-port's stability requires that the Γ_S-plane stability circle also be studied. The Γ_S-plane stability circle corresponds to the values of Γ_S which map to output reflection coefficients (Γ_{out}) with magnitude equal to one. The Γ_S-plane stability circle is described by

$$(U_S - U_{CS})^2 + (V_S - V_{CS})^2 = {r_S}^2 \tag{8.75}$$

where

$$C_S = U_{CS} + jV_{CS} \tag{8.76}$$

$$= \frac{S_{11}^* - D^* S_{22}}{|S_{11}|^2 - |D|^2} \tag{8.77}$$

$$r_S = \frac{|S_{12}S_{21}|}{||S_{11}|^2 - |D|^2|} \tag{8.78}$$

A decision regarding whether the unstable region in the Γ_S-plane lies inside or outside of the Γ_S-plane stability circle is made by examining the magnitude of S_{22} and noting that the point $\Gamma_S = 0$ (the origin of the Γ_S-plane) maps to S_{22}.

It is now appropriate to define the concept of unconditional stability. An *unconditionally stable* 2-port has the property that no choice of *passive* source and load terminations will make the 2-port oscillate. In other words, an unconditionally stable 2-port has $|\Gamma_{in}| < 1$ and $|\Gamma_{out}| < 1$ for any choice of passive source and load terminations. The restriction to passive sources and loads means that we limit our attention to sources with $|\Gamma_S| \leq 1$ and loads with $|\Gamma_L| \leq 1$, i.e., those regions of the Γ_S- and Γ_L-planes that lie inside and on the unit circle centered on the origin. A 2-port is unconditionally stable if the unstable regions of the Γ_S- and Γ_L-planes lie completely outside of the unit circles. Once the center coordinates and the radii of the stability circles have been computed, the circles can be plotted and, utilizing the known magnitudes of S_{11} and S_{22}, we can determine whether the unstable regions lie outside of the unit circles.

As an alternative to plotting the stability circles, it is possible to derive a relatively simple algebraic criterion (or set of criteria) that must be satisfied in order for a 2-port to be unconditionally stable. The derivation of one such criterion is given in section 8.3.4. A summary of the various sets of criteria that have been derived is given here. In each case, if the criterion (or set of criteria) is satisfied, then the 2-port is unconditionally stable:

$$K > 1, \quad |S_{12}S_{21}| < 1 - |S_{11}|^2 \tag{8.79}$$

$$K > 1, \quad |S_{12}S_{21}| < 1 - |S_{22}|^2 \tag{8.80}$$

$$K > 1, \quad B_1 > 0 \tag{8.81}$$

$$K > 1, \quad B_2 > 0 \tag{8.82}$$

$$K > 1, \quad |D| < 1 \tag{8.83}$$

$$\mu_{ES} = \frac{1 - |S_{11}|^2}{|S_{22} - S_{11}^* D| + |S_{12}S_{21}|} > 1 \tag{8.84}$$

$$\mu'_{ES} = \frac{1 - |S_{22}|^2}{|S_{11} - S_{22}^* D| + |S_{12}S_{21}|} > 1 \tag{8.85}$$

where

$$K = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |D|^2}{2\,|S_{12}|\,|S_{21}|} > 1 \tag{8.86}$$

$$D = S_{11}\,S_{22} - S_{12}\,S_{21} \tag{8.87}$$

$$B_1 = 1 + |S_{11}|^2 - |D|^2 - |S_{22}|^2 \tag{8.88}$$

$$B_2 = 1 + |S_{22}|^2 - |D|^2 - |S_{11}|^2 \tag{8.89}$$

Any one set of criteria given by 8.79-8.85 is a necessary and sufficient set of criteria for a 2-port to be unconditionally stable. If any one of the sets of stability criteria are satisfied, then no choice of passive source and load terminations will make the 2-port oscillate. This statement applies only at the frequency where the S-parameters were measured. It is possible for a 2-port to be unconditionally stable in a certain frequency band but potentially unstable in some other frequency band. In practice, stability should be checked at many frequencies in the bandwidth within which the 2-port has appreciable gain. This requires that the S-parameters be measured at many frequencies. Checking one of the sets of conditions given in equations 8.79-8.85 is easier than plotting the stability circles, especially when one wishes to quickly determine the stability status over a wide frequency range; however if the 2-port turns out to be potentially unstable, then the stability circles provide detailed information on which terminations make the 2-port unstable.

Finally, note that the parameter "K" defined in equation 8.86 and used in criteria 8.79 through 8.83 is called the *Rollet Stability Factor*. While the condition K>1, by itself, does not guarantee that a 2-port is unconditionally stable, it turns out that K plays an important role in determining whether or not a 2-port can be simultaneously matched with passive terminations at both the input and output ports.

8.3.3 Example - 2-port stability analysis.

A 2-port has S-parameters ($Z_o = 50\,\Omega$)

$$S_{11} = 0.4\angle -20^\circ \tag{8.90}$$

$$S_{12} = 0.1\angle 40^\circ \tag{8.91}$$

$$S_{21} = 7.5\angle 150^\circ \tag{8.92}$$

$$S_{22} = 0.6\angle -30^\circ \tag{8.93}$$

The Rollett Stability factor for this 2-port is $K = 0.853 < 1$, so the 2-port is potentially unstable. The coordinates of the stability circles are

$$\begin{array}{lllllll} U_{CS} = & 0.163 & V_{CS} = & -0.589 & r_S = & 1.17 \\ U_{CL} = & -0.401 & V_{CL} = & -0.913 & r_L = & 1.70 \end{array} \tag{8.94}$$

The stability circles for this 2-port are plotted in Figures 8.10 and 8.11. The stable region of the Γ_S- and Γ_L-planes is shaded. For the Γ_S-plane stability circle, the stable region was identified by noting that the origin of the Γ_S-plane ($\Gamma_S = 0$) maps to an output reflection coefficient $\Gamma_{out} = S_{22}$. Since the origin of the Γ_S-plane is in the interior of the stability circle and $|S_{22}| < 1$, we conclude that the interior of the stability circle maps to output reflection coefficients $|\Gamma_{out}| < 1$; hence, the interior of the stability circle is the stable region. Similar reasoning was used to determine that the interior of the Γ_L-plane stability circle was stable. To verify that the stable and unstable regions have been correctly identified, consider the

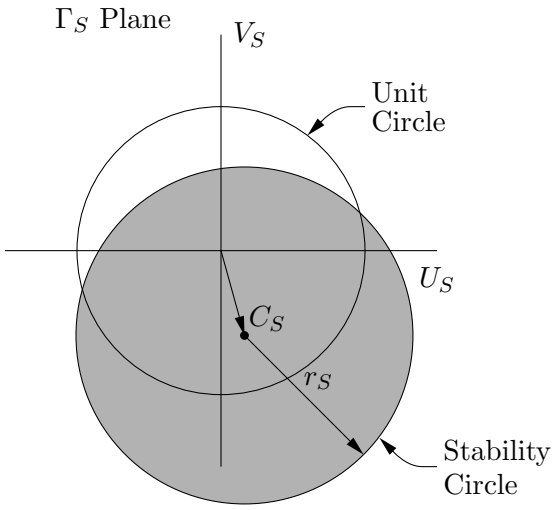


Figure 8.10: Γ_S plane stability circle. The stable region, corresponding to values of $|\Gamma_S|$ that map to $|\Gamma_{out}| < 1$, is shaded.

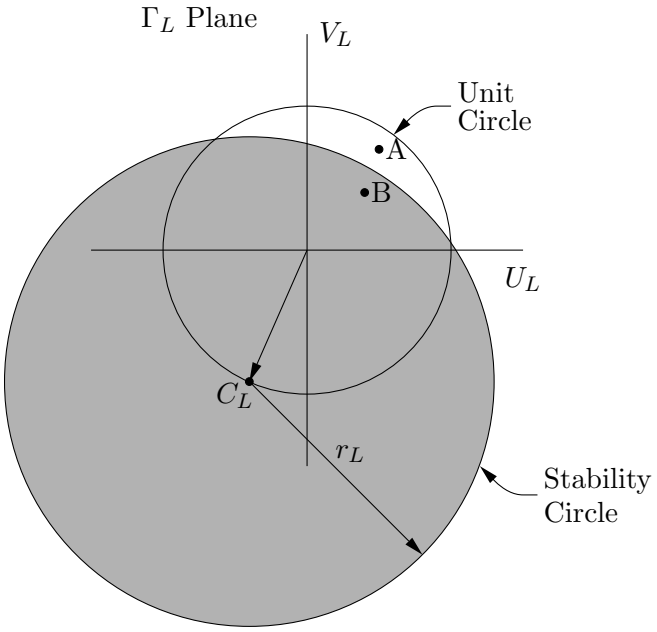


Figure 8.11: Γ_L plane stability circle. The stable region, corresponding to values of Γ_L that map to $|\Gamma_{in}| < 1$, is shaded. Point "A", at $\Gamma_L = 0.5 + j0.7 = 0.86\angle 54°$, lies in the unstable region. If the 2-port is terminated with this load reflection coefficient, the input reflection coefficient will have magnitude greater than one, corresponding to an input impedance with negative real part ($\Re(Z_{in}) < 0$).

two points labeled A and B in the Γ_L-plane plot. The point A is in the unstable region and represents a load reflection coefficient of $\Gamma_L = 0.5 + j0.7 = 0.86\angle 54°$. Using this value in Equation 8.48 yields $\Gamma_{in} = 1.3\angle -76.4°$ which represents an input impedance with a negative real part, as expected. Point B, which is inside the stable region, represents a load reflection coefficient $\Gamma_L = 0.4 + j0.4 = .57\angle 45°$. The input reflection coefficient is $\Gamma_{in} = 0.70\angle -83°$.

The following example is identical to the previous one, except that the magnitude of S_{12} has been reduced from 0.1 to 0.01.

2-port stability analysis. A 2-port has S-parameters ($Z_o = 50\,\Omega$):

$$S_{11} = 0.4\angle -20° \tag{8.95}$$

$$S_{12} = 0.01\angle 40° \tag{8.96}$$

$$S_{21} = 7.5\angle 150° \tag{8.97}$$

$$S_{22} = 0.6\angle -30° \tag{8.98}$$

The stability factor, K, is 3.74 (> 1), and we also have

$$1 - |S_{11}|^2 > |S_{12}S_{21}| \tag{8.99}$$

$$1 - |S_{22}|^2 > |S_{12}S_{21}| \tag{8.100}$$

so the 2-port is unconditionally stable. (Note, only one of 8.99 and 8.100 needs to be checked.) Stability circles for this example are shown in Figures 8.12 and 8.13. As in the first example, we have made use of the fact that the origin of the Γ_S- and Γ_L-planes map to output and input reflection coefficients, respectively, with magnitudes less than 1. Hence the origin of each plane is in the stable region. Since the origin is outside the stability circle, the region outside the circle is the stable region and the region inside is the unstable region. Since the unstable regions do not include any passive source or load terminations, the 2-port is said to be unconditionally stable.

In the following section, a simple test is derived that can predict whether or not a particular 2-port is unconditionally stable.

8.3.4 Derivation of a Criterion for Unconditional Stability of 2-ports

All of the essential information necessary to investigate the stability of a 2-port (at a particular frequency) is contained in the 4 parameters that define the stability circles in the Γ_S and Γ_L planes. In this section, we derive a criterion that can be used to quickly check to see if a 2-port is unconditionally stable. As already pointed out, numerous sets of necessary and sufficient criteria have been derived that can be used to determine whether or not a 2-port is unconditionally stable.[1] The criterion derived in this section is particularly interesting because it involves only a single inequality and because the numerical value of

[1]Lombardi, G. and B. Neri, Criteria for the evaluation of unconditional stability of microwave linear two-ports: a critical review and new proof, IEEE Trans. MTT, Vol. 47, No. 6, June, 1999, p746.

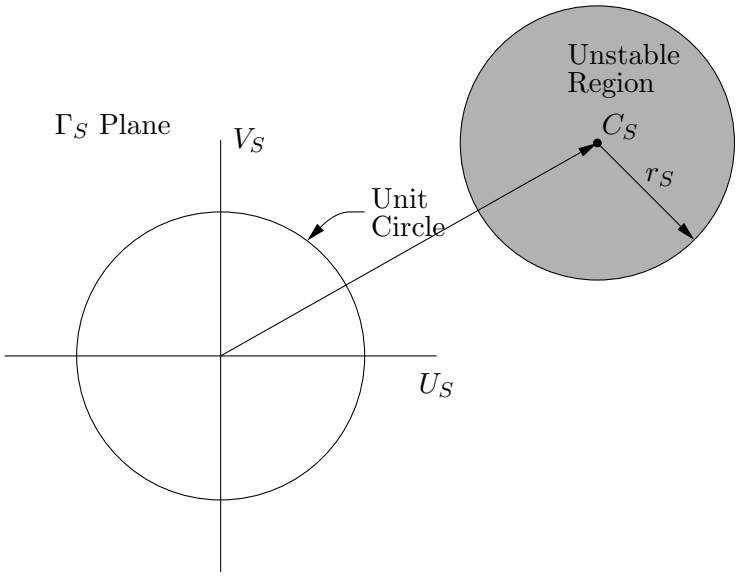


Figure 8.12: Γ_S plane stability circle. The unstable region is shaded.

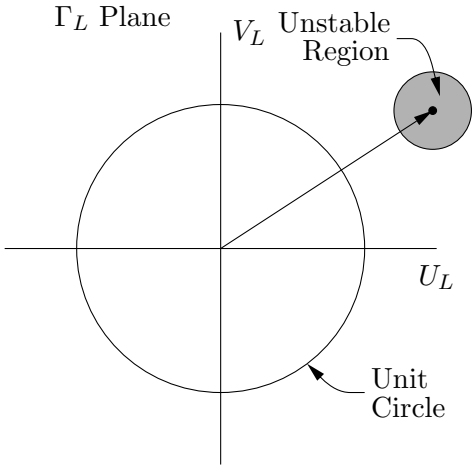


Figure 8.13: Γ_L plane stability circle. The unstable region is shaded.

the stability parameter has the useful property that its magnitude represents the smallest distance between the origin of the reflection coefficient plane and the edge of the unstable region.

In order for a 2-port to be unconditionally stable, it is necessary for the unstable region in the Γ_L and Γ_S planes to lie outside of the unit circle. Consider the Γ_L plane first, and assume that $|S_{11}| < 1$ (this is a necessary condition for unconditional stability, because if $|S_{11}| \geq 1$ a negative, or zero, real part of Z_{IN} is obtained by setting $Z_L = Z_o$. In this case, the 2-port can oscillate with a passive source termination and hence is potentially unstable.)

There are two cases to consider.

1. First, suppose that the stability circle does not enclose the origin of the Γ_L plane (i.e. $|C_L| > r_L$). Then the inside of the stability circle is the unstable region. If the 2-port is unconditionally stable it is necessary for the minimum distance between the origin and the edge of the stability circle to be larger than 1, i.e. it is necessary that $|C_L| - r_L > 1$. Inserting the equations for $|C_L|$ and r_L we find that this amounts to:

$$\frac{|S_{22}^* - D^* S_{11}|}{||S_{22}|^2 - |D|^2|} - \frac{|S_{12}S_{21}|}{||S_{22}|^2 - |D|^2|} > 1 \tag{8.101}$$

or

$$\frac{|S_{22}^* - D^* S_{11}| - |S_{12}S_{21}|}{||S_{22}|^2 - |D|^2|} > 1 \tag{8.102}$$

2. Now consider the other possibility, namely that the stability circle encloses the origin, i.e. that $|C_L| < r_L$. In this case, the necessary condition $|S_{11}| < 1$ implies that the exterior of the stability circle is the unstable region. In order to have the unstable region lie completely outside of the unit circle, we require that the stability circle completely encloses the unit circle centered at $\Gamma_L = 0$, i.e.:

$$r_L - |C_L| > 1$$

or

$$\frac{|S_{12}S_{21}| - |S_{22}^* - DS_{11}^*|}{||S_{22}|^2 - |D|^2|} > 1 \tag{8.103}$$

Note carefully that equation 8.102 guarantees that the unstable region is outside of the unit circle provided that $|C_L| > r_L$ whereas equation 8.103 guarantees that the unstable region is outside of the unit circle when $|C_L| < r_L$. Now consider what the statement $|C_L| > r_L$ (or $|C_L| < r_L$) implies in terms of the S parameters. Consider $|C_L| > r_L$:

$$\frac{|S_{22}^* - D^* S_{11}|}{||S_{22}|^2 - |D|^2|} > \frac{|S_{12}S_{21}|}{||S_{22}|^2 - |D|^2|}$$

or

$$|S_{22}^* - D^* S_{11}| > |S_{12}S_{21}|$$

Squaring both sides (note that squaring both sides does not lose any generality, because both sides are always positive) yields

$$(S_{22}^* - D^* S_{11})(S_{22} - DS_{11}^*) > |S_{12}S_{21}|^2$$

and expanding the LHS yields

$$|S_{22}|^2 - S_{22}^* D S_{11}^* - S_{22} D^* S_{11} + |D|^2 |S_{11}|^2 > |S_{12}S_{21}|^2$$

Expand the second and third terms on the LHS using $D = S_{11}S_{22} - S_{12}S_{21}$:

$$|S_{22}|^2 - S_{11}^* S_{22}^* (S_{11}S_{22} - S_{12}S_{21}) - S_{11}S_{22}(S_{11}^* S_{22}^* - S_{12}^* S_{21}^*) + |D|^2|S_{11}|^2 > |S_{12}S_{21}|^2$$

and group terms on the LHS:

$$|S_{22}|^2 - 2|S_{11}|^2|S_{22}|^2 + S_{11}^* S_{22}^* S_{12}S_{21} + S_{11}S_{22}S_{12}^* S_{21}^* + |D|^2|S_{11}|^2 > |S_{12}S_{21}|^2 \quad (8.104)$$

Now, noting that

$$|D|^2 = |S_{11}|^2|S_{22}|^2 - S_{11}S_{22}S_{12}^* S_{21}^* - S_{11}^* S_{22}^* S_{12}S_{21} + |S_{12}|^2|S_{21}|^2 \quad (8.105)$$

Equation 8.105 can be used to eliminate the complex terms (3rd and 4th term on LHS) in equation 8.104. The result is:

$$|S_{22}|^2 - |S_{11}|^2|S_{22}|^2 + |S_{12}|^2|S_{21}|^2 - |D|^2 + |D|^2|S_{11}|^2 > |S_{12}|^2|S_{21}|^2$$

or, after grouping terms, we determine that $|C_L| > r_L$ amounts to:

$$(|S_{22}|^2 - |D|^2)(1 - |S_{11}|^2) > 0 \quad (8.106)$$

Since we have already assumed that $|S_{11}| < 1$, the second term on the LHS is always positive so the combined requirements that the stability circle not enclose the origin ($|C_L| > r_L$) and that the inside of the stability circle represent the unstable region ($|S_{11}| < 1$) means that:

$$|S_{22}|^2 - |D|^2 > 0 \quad (8.107)$$

Similarly, the requirements that the stability circle encloses the origin ($|C_L| < r_L$) and that the inside of the stability circle represents the unstable region ($|S_{11}| < 1$) amounts to:

$$|S_{22}|^2 - |D|^2 < 0 \quad (8.108)$$

We now know that criterion 8.79 applies only when $|S_{22}|^2 - |D|^2 > 0$ and criterion 8.80 applies when $|S_{22}|^2 - |D|^2 < 0$. Note that criteria 8.79 and 8.80 differ only in the sign of the numerator on the LHS, and that the denominator is the magnitude of $|S_{22}|^2 - |D|^2$. Thus, these criteria can be combined into into a single inequality which is valid for either case:

$$\frac{|S_{22}^* - D^* S_{11}| - |S_{12}S_{21}|}{|S_{22}|^2 - |D|^2} > 1 \quad (8.109)$$

There is an apparent singularity in equation 8.109 when $|S_{22}|^2 - |D|^2 = 0$. This corresponds to the situation where the radius of the stability circle approaches infinity, i.e. when the stability circle degenerates into a straight line. The singularity can be removed by noting that:

$$|S_{22}|^2 - |D|^2 = \frac{|S_{22} - S_{11}^* D|^2 - |S_{12}S_{21}|^2}{1 - |S_{11}|^2}$$
$$= \frac{(|S_{22} - S_{11}^* D| - |S_{12}S_{21}|)(|S_{22} - S_{11}^* D| + |S_{12}S_{21}|)}{1 - |S_{11}|^2} \quad (8.110)$$

Use equation 8.110 in equation 8.109 to write:

$$\mu_{ES} = \frac{1 - |S_{11}|^2}{|S_{22}^* - D^* S_{11}| + |S_{12}S_{21}|} > 1 \quad (8.111)$$

The parameter μ_{ES} is referred to as the Edwards-Sinsky stability criterion,[2]and has a useful geometric interpretation: μ_{ES} is the minimum distance between the origin of the Γ_L plane and the unstable region. If the parameter μ_{ES} is negative, it means that the unstable region includes the origin of the Γ_L plane.

If the identical analysis is carried out for the Γ_S plane stability circles, we derive the dual constraint:

$$\mu_{ES}^{'} = \frac{1 - |S_{22}|^2}{|S_{11}^* - D^* S_{22}| + |S_{12}S_{21}|} > 1 \tag{8.112}$$

The geometric interpretation of $\mu_{ES}^{'}$ is the same as that of μ_{ES}, but applied to the Γ_S plane.

It turns out that either criterion 8.111 or criterion 8.112 is a necessary and sufficient test for unconditional stability of a 2-port, even though each one was derived separately by considering only one of the Γ_L plane or Γ_S plane stability circles. To show that this is true we can prove that "if $\mu_{ES} \leq 1$ then $\mu_{ES}^{'} \leq 1$" and the converse are both true statements. Suppose $\mu_{ES} \leq 1$ and choose a passive load termination, call it Γ_{Lu}, that causes $|\Gamma_{in}|$ to be ≥ 1 , i.e. $|\Gamma_{in}(\Gamma_{Lu})| \geq 1$. The negative resistance criterion for steady-state oscillation will be satisfied at the input port if the source reflection coefficient is chosen to be $\Gamma_{Su} = \Gamma_{in}(\Gamma_{Lu})^{-1}$ because this choice causes $\Gamma_{Su}\Gamma_{in}(\Gamma_{Lu}) = 1$. (As an exercise, you may wish to verify that the condition $Z_1 + Z_2 = 0$ is equivalent to $\Gamma_1\Gamma_2 = 1$.) Note that Γ_{Su} is passive, since $|\Gamma_{Su}| = 1/|\Gamma_{in}(\Gamma_{Lu})| \leq 1$. Now, it is left as an exercise to show that $\Gamma_{out}(\Gamma_{Su}) = \Gamma_{Lu}^{-1}$. Thus $\Gamma_{out}(\Gamma_{Su})\Gamma_{Lu} = 1$, which means that the oscillation condition is also satisfied at the output port. The converse can be proven by simply changing subscripts in the preceeding argument. It follows that if $\mu_{ES} > 1$ then $\mu_{ES}^{'} > 1$ and vice versa. Hence we can determine whether or not a 2-port is unconditionally stable by checking either criterion 8.111 or 8.112.

8.3.5 Terminations for Simultaneous Conjugate Match

Whenever $K > 1$ it is possible to find a combination of passive source and load terminations for which both the input and output of the 2-port are conjugately matched. For a given available source power, a simultaneous conjugate match condition leads to the highest possible power delivered to the load. The particular source and load reflection coefficients which result in a simultaneous conjugate match at both ports are denoted by Γ_{ms} and Γ_{ml}. Equations for Γ_{ms} and Γ_{ml} can be derived by solving Equations 8.113 and 8.114, which enforce the conjugate match relationship at the input and output of the 2-port, i.e.,

$$\Gamma_{in} = \Gamma_{ms}^* = S_{11} + \frac{S_{12}S_{21}\Gamma_{ml}}{1 - S_{22}\Gamma_{ml}} \tag{8.113}$$

$$\Gamma_{out} = \Gamma_{ml}^* = S_{22} + \frac{S_{12}S_{21}\Gamma_{ms}}{1 - S_{11}\Gamma_{ms}} \tag{8.114}$$

The solution to Equations 8.113 and 8.114 is given by

$$\begin{aligned} \Gamma_{ms} &= \frac{B_1 \pm \sqrt{B_1^2 - 4|C_1|^2}}{2C_1} \\ B_1 &= 1 + |S_{11}|^2 - |D|^2 - |S_{22}|^2 \\ C_1 &= S_{11} - DS_{22}^* \end{aligned} \tag{8.115}$$

[2]Edwards, M. L. and J. H. Sinsky, A new criterion for linear 2-port stability using a single geometrically derived parameter, IEEE Trans. MTT, Vol. 40., No. 12, December 1992, p2303.

$$\begin{aligned} \Gamma_{ml} &= \frac{B_2 \pm \sqrt{B_2^2 - 4|C_2|^2}}{2C_2} \\ B_2 &= 1 + |S_{22}|^2 - |D|^2 - |S_{11}|^2 \\ C_2 &= S_{22} - DS_{11}^* \end{aligned} \tag{8.116}$$

The proper choice of sign in Equations 8.115 and 8.116 is determined by the requirement that $|\Gamma_{ms}| < 1$ and $|\Gamma_{ml}| < 1$. In other words, the correct solution in each case is the one with a magnitude less than 1. Equations 8.115 and 8.116 will yield valid solutions whenever $K > 1$. Since this is only a necessary condition for stability, it is possible to find a simultaneous conjugate match solution for some potentially unstable 2-ports (those with $K > 1$).

Suppose Γ_{ms} and Γ_{ml} are known for a particular 2-port (with $K > 1$). Then the procedure for designing an amplifier which maximizes the power delivered to the load consists of designing matching networks that transform the actual source and load reflection coefficients into Γ_{ms} and Γ_{ml}, as in Figure 8.14.

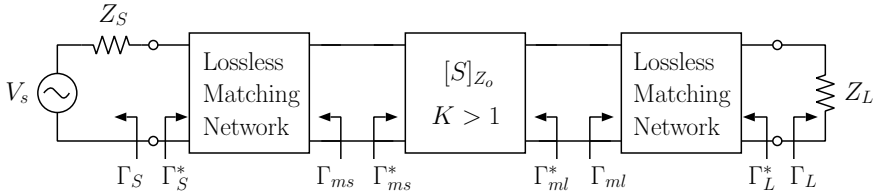


Figure 8.14: An amplifier that is simultaneously matched to source and load terminations. A conjugate match exists at all ports.

If a 2-port does not satisfy the conditions for unconditional stability, then it is said to be potentially unstable. Useful and stable amplifiers can be designed with potentially unstable 2-ports, however it is necessary to carefully choose source and load terminations to avoid the unstable regions of the Γ_S- and Γ_L- planes. If $K > 1$, then one such choice of terminations is $\Gamma_S = \Gamma_{ms}$ and $\Gamma_L = \Gamma_{ml}$. If $K \leq 1$, then it will not be possible to achieve a simultaneous conjugate match, and the terminations will have to be chosen according to a carefully measured tradeoff between transducer gain and relative stability. A 2-port with $K \leq 1$ can be modified using resistive loading to raise the stability factor so that the loaded 2-port can be simultaneously conjugate-matched.

Of course, potentially unstable 2-ports can also be used to design oscillators. For example, a basic oscillator design could be accomplished by choosing a Γ_S in the unstable region, i.e., a Γ_S that maps to $|\Gamma_{out}| > 1$. Using this source termination, the output impedance of the 2-port will have a negative real part. The load termination is then chosen such that the negative resistance criterion for oscillation is satisfied at the output port. As shown earlier, this will automatically cause the negative resistance criterion to be satisfied at the input port.

8.3.6 Power Gains

1. **Operating Power Gain** (or just "power gain")

$$G \equiv \frac{P_{out}}{P_{in}} = \frac{|S_{21}|^2\left(1-|\Gamma_L|^2\right)}{(1-|S_{11}|^2)+|\Gamma_L|^2(|S_{22}|^2-|D|^2)-2Re(\Gamma_L C_2)} \tag{8.117}$$

with

$$D = S_{11}S_{22} - S_{12}S_{21}$$

$$C_2 = S_{22} - DS_{11}^*$$

Note that G depends only on the load termination, Γ_L.

2. **Transducer Power Gain**

$$G_T \equiv \frac{P_{out}}{P_{avs}} = \frac{|S_{21}|^2\left(1-|\Gamma_S|^2\right)\left(1-|\Gamma_L|^2\right)}{|(1-S_{11}\Gamma_S)(1-S_{22}\Gamma_L)-S_{12}S_{21}\Gamma_L\Gamma_S|^2} \tag{8.118}$$

Notice that G_T depends on both terminations, Γ_S and Γ_L.

3. **Available Power Gain**

$$G_A \equiv \frac{P_{avo}}{P_{avs}} = \frac{|S_{21}|^2\left(1-|\Gamma_S|^2\right)}{(1-|S_{22}|^2)+|\Gamma_S|^2(|S_{11}|^2-|D|^2)-2Re(\Gamma_S C_1)} \tag{8.119}$$

with

$$D = S_{11}S_{22} - S_{12}S_{21}$$

$$C_1 = S_{11} - DS_{22}^*$$

Note that G_A depends solely on the source impedance. It tells us how much power is potentially available from the output of the 2-port.

4. **Unilateral Transducer Power Gain**

If the internal feedback within the device is small ($S_{12} \approx 0$), then the device can be considered unilateral. This approximation yields a particularly simple form for the transducer gain:

$$\begin{aligned} G_{TU} &= G_T|_{S_{12}=0} \\ G_{TU} &= \frac{1-|\Gamma_S|^2}{|1-S_{11}\Gamma_S|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \end{aligned} \tag{8.120}$$

5. **Maximum Available Gain**

A 2-port that is conjugately matched at both ports (as in Figure 8.14) will have $G = G_A = G_T = G_{A,max}$ where:

$$G_{A,max} = \left|\frac{S_{21}}{S_{12}}[K \pm \sqrt{K^2-1}]\right| \tag{8.121}$$

The upper sign applies when $B_1 < 0$ and the lower sign applies when $B_1 > 0$. The maximum available gain could also be calculated by using Γ_{ml} in the formula for G, or Γ_{ms} in the formula for G_A, or (Γ_{ms}, Γ_{ml}) in the formula for G_T.

6. **Maximum Stable Gain**

   A 2-port with $K < 1$ cannot be simultaneously conjugate-matched at both ports using passive terminations. The stability factor of such a 2-port can be increased using resistive loading at the input and/or output. If the resistive loading is chosen to raise the stability factor of the loaded 2-port to a value slightly larger than 1, then the loaded 2-port can be simultaneously conjugate-matched at both ports and will have $G_{A,max} \simeq |\frac{S_{21}}{S_{12}}|$. Hence, for a 2-port with $K < 1$ we define the maximum stable gain to be:

$$G_{MS} = \left|\frac{S_{21}}{S_{12}}\right| \tag{8.122}$$

8.3.7 Example - Mismatch factor in terms of Γ_S and Γ_L

Consider a load with reflection coefficient Γ_L connected directly to a source with reflection coefficient Γ_S. We can obtain an expression for the mismatch factor, MF, by considering the load to be coupled to the source through a 2-port consisting of zero-length wires connecting the input and output terminals. Such a 2-port has the following S-parameter matrix:

$$[S]_{Z_o} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{8.123}$$

Using this S-matrix in the transducer gain expression (equation 8.118) yields the expression for the mismatch factor:

$$MF = \frac{(1-|\Gamma_S|^2)(1-|\Gamma_L|^2)}{|1-\Gamma_L\Gamma_S|^2} \tag{8.124}$$

8.3.8 Example - Power transfer to an antenna through a lossy transmission line

A lossy transmission line can be modeled as a 2-port with the following S-parameter matrix:

$$[S]_{Z_o} = \begin{bmatrix} 0 & e^{-\gamma l} \\ e^{-\gamma l} & 0 \end{bmatrix} \tag{8.125}$$

where the reference impedance for the S-parameter matrix, Z_o, is assumed to be equal to the characteristic impedance of the transmission line. The length of the transmission line is denoted by l and the complex propagation constant on the line is $\gamma = \alpha + j\beta$.

Suppose that a lossy transmission line (characteristic impedance Z_o) is used to connect a source with impedance Z_o to an antenna with impedance Z_L (and corresponding reflection coefficient Γ_L). Consider three cases:

1. The transmitter is connected to the input of the transmission line and the antenna is connected to the output of the line. No matching networks are used. In this case the power delivered to the load is:

$$P_L = P_{avs}G_T = P_{avs}e^{-2\alpha l}(1-|\Gamma_L|^2) \tag{8.126}$$

2. The transmitter is coupled to the input of the line through a lossless matching network whereas the antenna is connected directly to the output of the lossy line. In this case the power delivered to the load is

$$P_L = P_{avs}G = P_{avs}e^{-2\alpha l}\frac{1-|\Gamma_L|^2}{1-e^{-4\alpha l}|\Gamma_L|^2} \tag{8.127}$$

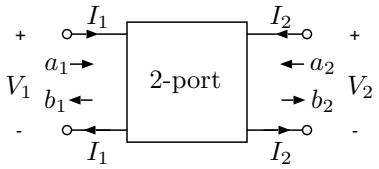
3. The transmitter is connected directly to the input of the line, and a lossless matching network is used in between the output of the lossy line and the antenna. In this case the power delivered to the load is

$$P_L = P_{avs} G_A = P_{avs} e^{-2\alpha l} \tag{8.128}$$

Notice that the power delivered to the antenna will be largest in case 3, i.e. when the antenna impedance is transformed to Z_o with a matching network located at the antenna[3]. In the limit $\alpha \to 0$, the line becomes lossless and cases 2 and 3 converge to the same result ($P_L = P_{avs}$). This is consistent with the fact that a passive lossless 2-port that is conjugately matched at one port is automatically matched at the other port. Therefore, in the lossless line case a single matching network at either the source or the load end of the line is sufficient to provide optimum power transfer to the load.

[3]The factor $e^{-2\alpha l}$ is the loss when the load impedance is equal to the characteristic impedance of the line. This is specified on datasheets for transmission lines in units of dB per unit length, i.e. the datasheet might specify the quantity $10\log(e^{2\alpha l})$ in dB for $l = 1$ m or $l = 100$ feet or some other standard length.

8.4 Summary of useful S-parameter formulas



Relationship between impedance Z_X (X can be *in*, *out*, S, L) and reflection coefficient Γ_X :

$$\Gamma_X = \frac{Z_X - Z_o}{Z_X + Z_o} \quad Z_X = Z_o \frac{1+\Gamma_X}{1-\Gamma_X}$$

Input reflection coefficient with arbitrary Z_L :

$$\Gamma_{in} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L}$$

Output reflection coefficient with arbitrary Z_S :

$$\Gamma_{out} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S}$$

Voltage gain with arbitrary Z_L (voltage gain does not depend on Z_S):

$$A_V = \frac{V_2}{V_1} = \frac{S_{21}(1+\Gamma_L)}{(1+S_{11})(1-S_{22}\Gamma_L) + S_{12}S_{21}\Gamma_L}$$

Current gain with arbitrary Z_L (current gain does not depend on Z_S):

$$A_I = \frac{I_2}{I_1} = \frac{S_{21}(\Gamma_L - 1)}{(1-S_{11})(1-S_{22}\Gamma_L) - S_{12}S_{21}\Gamma_L}$$

Definitions of some commonly used quantities:

$$\begin{aligned}
D &= S_{11}S_{22} - S_{12}S_{21} \\
K &= \frac{1-|S_{11}|^2-|S_{22}|^2+|D|^2}{2|S_{12}S_{21}|} \\
B_1 &= 1 + |S_{11}|^2 - |D|^2 - |S_{22}|^2 \\
C_1 &= S_{11} - DS_{22}^* \\
B_2 &= 1 + |S_{22}|^2 - |D|^2 - |S_{11}|^2 \\
C_2 &= S_{22} - DS_{11}^*
\end{aligned}$$

Operating Power Gain:

$$G \equiv \frac{\mathrm{P_{out}}}{\mathrm{P_{in}}} = \frac{|S_{21}|^2\left(1-|\Gamma_L|^2\right)}{(1-|S_{11}|^2) + |\Gamma_L|^2(|S_{22}|^2 - |D|^2) - 2Re(\Gamma_L C_2)}$$

Available Power Gain:

$$G_A \equiv \frac{\mathrm{P_{avo}}}{\mathrm{P_{avs}}} = \frac{|S_{21}|^2\left(1-|\Gamma_S|^2\right)}{(1-|S_{22}|^2) + |\Gamma_S|^2(|S_{11}|^2 - |D|^2) - 2Re(\Gamma_S C_1)}$$

Transducer Power Gain:

$$G_T \equiv \frac{\mathrm{P_{out}}}{\mathrm{P_{avs}}} = \frac{|S_{21}|^2\left(1-|\Gamma_S|^2\right)\left(1-|\Gamma_L|^2\right)}{|(1-S_{11}\Gamma_S)(1-S_{22}\Gamma_L) - S_{12}S_{21}\Gamma_L\Gamma_S|^2}$$

A 2-port is unconditionally stable if:

$$K > 1 \text{ and } 1 - |S_{11}|^2 > |S_{12}S_{21}|$$

or, if:

$$K > 1 \text{ and } 1 - |S_{22}|^2 > |S_{12}S_{21}|$$

or, if:

$$K > 1 \text{ and } B_1 > 0$$

or, if:

$$K > 1 \text{ and } B_2 > 0$$

or, if:

$$K > 1 \text{ and } |D| < 1$$

or, if:

$$\mu_{ES} = \frac{1 - |S_{11}|^2}{|S_{22} - S_{11}^* D| + |S_{12}S_{21}|} > 1$$

or, if:

$$\mu'_{ES} = \frac{1 - |S_{22}|^2}{|S_{11} - S_{22}^* D| + |S_{12}S_{21}|} > 1$$

Stability circles in the $\Gamma_L -$ plane:

$$\text{Center is at : } \quad C_L = \frac{S_{22}^* - D^* S_{11}}{|S_{22}|^2 - |D|^2} \qquad \text{radius : } \quad r_L = \frac{|S_{12}S_{21}|}{||S_{22}|^2 - |D|^2|}$$

Stability circles in the $\Gamma_S -$ plane:

$$\text{Center is at : } \quad C_S = \frac{S_{11}^* - D^* S_{22}}{|S_{11}|^2 - |D|^2} \qquad \text{radius : } \quad r_S = \frac{|S_{12}S_{21}|}{||S_{11}|^2 - |D|^2|}$$

Source and load reflection coefficient for simultaneous conjugate match (K must be > 1) - in each case, choose the sign that results in a reflection coefficient with magnitude less than 1:

$$\Gamma_{ms} = \frac{B_1 \pm \sqrt{B_1^2 - 4|C_1|^2}}{2C_1} \qquad \Gamma_{ml} = \frac{B_2 \pm \sqrt{B_2^2 - 4|C_2|^2}}{2C_2}$$

Maximum available power gain (only defined for $K > 1$):

$$G_{A,max} = \left| \frac{S_{21}}{S_{12}} [K \pm \sqrt{K^2 - 1}] \right|$$

$G_{A,max}$ is defined only for 2-ports that can be conjugately matched at both ports ($K > 1$). For unconditionally stable 2-ports, $B_1 > 0$, use the lower (negative) sign. For potentially unstable 2-ports, $B_1 \leq 0$, use the upper (positive) sign.

8.5 References

1. Carson, Ralph S., *High Frequency Amplifiers*, John Wiley & Sons, New York, 1975.

2. Gonzalez, Guillermo, *Microwave Transistor Amplifiers: Analysis and Design*, Prentice-Hall, New Jersey, 1984.

3. Liao, Samuel Y., *Microwave Circuit Analysis and Amplifier Design*, Prentice-Hall, New Jersey, 1987.

4. Pozar, David M., *Microwave Engineering*, Addison-Wesley Publishing Company, 1990.

5. Vendelin, George D., *Design of Amplifiers and Oscillators by the S-parameter Method*, John Wiley & Sons, New York, 1982.

6. Vendelin, George D., Anthony M. Pavio, Ulrich L. Rohde, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, John Wiley & Sons, 1990.

8.6 Homework Problems

1. Find the S-parameters for the T-network in Figure 8.15. Port 1 is on the left. Denote the reference impedance by Z_o.

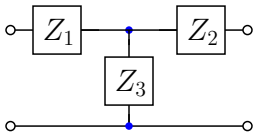


Figure 8.15: T-network

2. Find expressions for the S-parameters of the 2-port in Figure 8.16. Denote the reference impedance by Z_o.

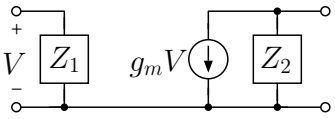


Figure 8.16: Unilateral hybrid-Pi model.

3. Consider an ideal transformer as shown in Figure 8.17. This 2-port is characterized

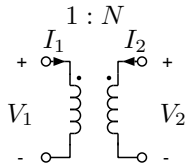


Figure 8.17: Ideal transformer

by the following relations:

$$\begin{aligned} V_2 &= NV_1 \\ I_2 &= -I_1/N \end{aligned} \tag{8.129}$$

Find expressions for the S-parameters of this 2-port. Denote the reference impedance by Z_o.

4. Suppose that two 2-ports are cascaded. Show that the S_{21} parameter for the overall network is given by

$$S_{21} = \frac{S_{21}^{(1)} S_{21}^{(2)}}{1 - S_{11}^{(2)} S_{22}^{(1)}} \tag{8.130}$$

where $S_{ij}^{(1)}$ and $S_{ij}^{(2)}$ refer to the S-parameters of the first and second 2-ports, respectively.

5. A 2-port has the following S-parameters ($Z_0 = 50\,\Omega$):

$$\begin{aligned} S_{11} &= 0.5\angle -96^\circ \\ S_{12} &= 0.3\angle 50^\circ \\ S_{21} &= 5.2\angle 45^\circ \\ S_{22} &= 0.4\angle -120^\circ \end{aligned} \tag{8.131}$$

   (a) What is the input impedance when the 2-port is terminated with a short circuit, i.e., $Z_L = 0$?

   (b) Is this 2-port unconditionally stable? (Check one of the necessary and sufficient sets of criteria.)

   (c) Find the coordinates of the center and the radius of the stability circles in the Γ_S and Γ_L planes. Sketch the stability circles and shade the regions that correspond to those values of Γ_S and Γ_L that make $|\Gamma_{out}| > 1$ or $|\Gamma_{in}| > 1$.

6. Suppose that S_{12} in Problem 5 is changed to $S_{12} = 0.05\angle 50^\circ$. Find the coordinates of the center and the radius of the stability circles in the Γ_S and Γ_L planes. Sketch the stability circles and shade the regions that correspond to those values of Γ_S and Γ_L that make $|\Gamma_{out}| > 1$ or $|\Gamma_{in}| > 1$. Is the 2-port unconditionally stable?

7. The stability circles in the output (Γ_L) plane for cases (a) and (b) are shown in Figure 8.18. Suppose that in both cases it is known that $|S_{11}| < 1$. For each case indicate

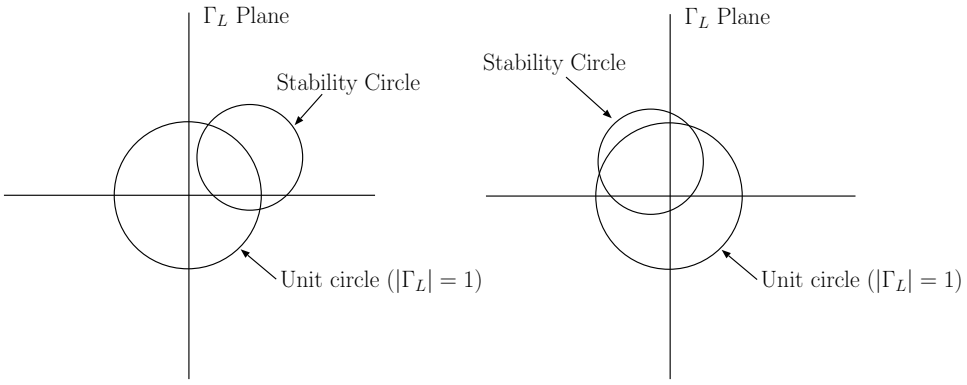


Figure 8.18: Case (a), left and case (b), right.

what region of the Γ_L plane corresponds to the stable region. The stable region of the Γ_L plane corresponds to those values of Γ_L that will cause the input reflection coefficient to have a magnitude less than 1.

8. For a particular 2-port with $|S_{11}| < 1$ and $|S_{22}| < 1$, the center and radius of the stability circles in the Γ_S and Γ_L planes are known to be

$$\begin{aligned} C_L &= -0.269 - j1.86 \\ r_L &= 3.01 \end{aligned} \tag{8.132}$$

$$\begin{aligned} C_S &= -2.23 + j4.76 \\ r_S &= 4.19 \end{aligned}$$

   (a) Make two sketches, one each for the Γ_S and Γ_L planes, and show the stability circles. Shade the unstable region.

   (b) Is the 2-port unconditionally stable?

9. Consider the system in Figure 8.19 where $R_S = 300\,\Omega$ and $R_L = 100\,\Omega$. Suppose that the 2-port is known to be unilateral and that $S_{22} = 0.8$ ($Z_o = 50\,\Omega$). The transducer gain of the 2-port is $G_T = 12$ dB. Find the available gain of the 2-port in this system.

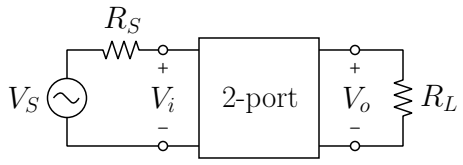


Figure 8.19: System with resistive source and load.

10. Consider the system in Figure 8.19 where $R_S = R_L = 300\,\Omega$. The voltage gain, $A_v = \frac{V_o}{V_i}$, in this system is found to be $A_v = 6\angle 45°$. In addition, it is known that the 2-port is unilateral, and that $S_{11} = 0.2\,(Z_o = 50\,\Omega)$.

   (a) Find the operating power gain. Express your result in dB.

   (b) Find the transducer powergain. Express your result in dB.

11. A 2-port has the following S-parameters ($Z_0 = 50\,\Omega$):

$$\begin{aligned} S_{11} &= 0.6\angle -100° \\ S_{12} &= 0.05\angle 45° \\ S_{21} &= 3.50\angle 60° \\ S_{22} &= 0.1\angle -30° \end{aligned} \tag{8.133}$$

   (a) Is this 2-port unconditionally stable?

   (b) Can this 2-port be conjugately matched simultaneously at the input and output? If so, find the source and load impedances (Z_{ms} and Z_{ml}) that will make this condition occur.

(c) If this 2-port is used in a system with the source and load impedances found in part 11b, find the operating, transducer, and available gains (G, G_T, and G_A). Express your results in dB.

12. The 2-port from Problem 11 is used in a system with open circuit source voltage $V_S = 3$V, source impedance $R_S = 100\Omega$, and load impedance 500Ω (see Figure 8.19).

(a) Compute the power available from the source, P_{avs}. Express your result in dBm.

(b) Find G, G_T, and G_A for the 2-port when used in this system. Express your results in dB

(c) Find the power delivered to the 2-port and the power delivered to the load. Express your result in dBm.

(d) How much power will be delivered to the load if a lossless matching network is added between the output of the 2-port and the 500Ω load? The matching network would be designed to transform 500Ω into Z_{out}^* where Z_{out}^* is the output impedance of the 2-port with the 100Ω source termination. Express your result in dBm.

13. Consider the 2-port from Problem 11. Suppose the reference impedance, Z_o, is changed to 500Ω. Denote the new S-parameters for this reference impedance by S'_{11}, S'_{21}, S'_{12} and S'_{22}.

(a) Find S'_{11}.

(b) Find S'_{21}. Hint: The voltage gain $A_v = V_o/V_i$ in a system with arbitrary load and source impedance is given by the following formula:

$$A_v = \frac{S_{21}(1+\Gamma_L)}{(1-S_{22}\Gamma_L)(1+\Gamma_{IN})} \tag{8.134}$$

where Γ_{IN} is the input reflection coefficient and V_o, V_i are the voltages measured across the output and input terminals of the 2-port, respectively.

14. A 2-port has 50 Ω S-parameters:

$$\begin{aligned} S_{11} &= 0.3 \\ S_{12} &= 0.01 \\ S_{21} &= 10.0 \\ S_{22} &= 0.1 \end{aligned} \tag{8.135}$$

The Rollett Stability Factor for this 2-port is K = 4.525, and the 2-port is unconditionally stable.

(a) The 2-port is used with a source having impedance $Z_s = 300\,\Omega$ and a load $Z_L = 50\,\Omega$. The power available from the source is -3 dBm. Find the power delivered to the load. Express your result in dBm.

(b) Now suppose a lossless matching network is used between the source and the 2-port. Find the power delivered to the load. Express your result in dBm.

(c) Suppose a lossless matching network is used between the 2-port and the load (no matching network at the input). Find the power delivered to the load. Express your result in dBm.

(d) Suppose both input and output matching networks are used so that both ports are conjugately matched. Find the power delivered to the load. Express your result in dBm.

15. Show that the maximum available gain $G_{A,max}$ for a **unilateral** 2-port is

$$G_{A,max} = \frac{|S_{21}|^2}{(1-|S_{11}|^2)(1-|S_{22}|^2)} \tag{8.136}$$

Hint: When the input and output of the 2-port are conjugately matched, $G = G_T = G_{A,max}$.

16. Consider a system with source and load reflection coefficients Γ_S and Γ_L as shown in Figure 8.20. Define the insertion gain

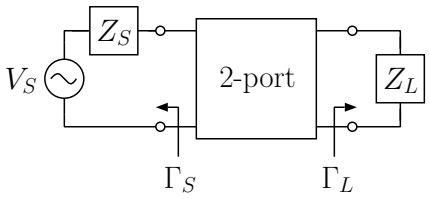


Figure 8.20: System with source and load reflection coefficients Γ_S and Γ_L

$$G_I = \frac{P_{\text{delivered to load with network}}}{P_{\text{delivered to load without network}}} \tag{8.137}$$

where $P_{\text{delivered to load without network}}$ is the power delivered to the load when the source is connected directly to the load, and $P_{\text{delivered to load with network}}$ is the power delivered to the load when the 2-port is used between the source and the load.

(a) Find an expression for the insertion gain of an arbitrary 2-port.

(b) Find an expression for the insertion gain of a lossless matching network.

17. The "T-parameters" (sometimes called the "chain scattering parameters") are defined in terms of the same variables as the S-parameters. They are useful when cascading 2-ports, because the T-parameter matrix for a cascade is simply the product of the T-matrices for the individual 2-ports. (Can you show that this is true?) The definition of the T-parameters follows from choosing a_1 and b_1 to be the dependent variables, i.e.,

$$\begin{aligned} a_1 &= T_{11}\, b_2 \,+\, T_{12}\, a_2 \\ b_1 &= T_{21}\, b_2 \,+\, T_{22}\, a_2 \end{aligned} \tag{8.138}$$

where the a's and b's are defined as

$$a_i = \frac{V_i + Z_o I_i}{2\sqrt{Z_o}} \tag{8.139}$$

$$b_i = \frac{V_i - Z_o I_i}{2\sqrt{Z_o}}. \tag{8.140}$$

Suppose the S-parameters for a 2-port are known. Find the T-parameters in terms of the S-parameters.

18. Consider a cascade of two identical 2-ports. The cascade is used in a system with source and load as shown in Figure 8.21:

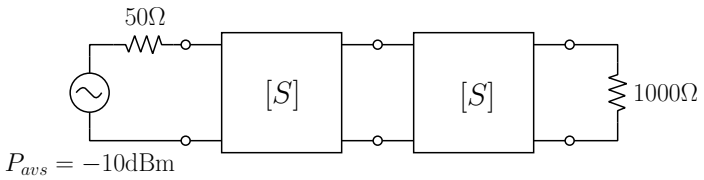


Figure 8.21: Cascade of two identical 2-ports

The available source power is -10dBm and the 50 Ω S-parameters of the 2-ports are

$$\begin{aligned} S_{11} &= 0.35 \\ S_{12} &= 0.1 \\ S_{21} &= 3.0 \\ S_{22} &= 0.50 \end{aligned} \tag{8.141}$$

The 2-ports are unconditionally stable. Answer the following questions. For parts 18a - 18g, express all results in dB or dBm, whichever is appropriate.

(a) The power delivered to the first 2-port.

(b) The power delivered to the second 2-port.

(c) The power delivered to the load.

(d) The transducer gain for the cascaded 2-ports.

(e) The available gain of the cascaded 2-ports.

(f) The operating power gain of the cascaded 2-ports.

(g) Suppose three lossless matching networks are used between the source and the first 2-port, the first and second 2-ports, and the second 2-port and the load. Find the power delivered to the load.

(h) Find S_{11} ($Z_o = 50$) for the cascaded 2-ports. The answer is NOT 0.35!

19. Consider the system shown in Figure 8.22.

The 2-port has the following S-parameters ($Z_o = 50\,\Omega$):

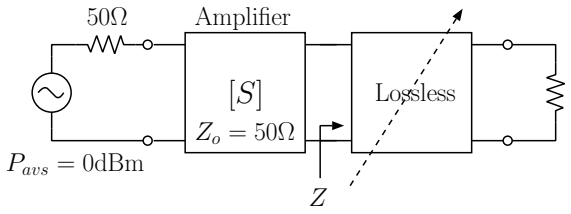


Figure 8.22: System with adjustable, lossless matching network between the 2-port and the load.

$$S_{11} = \quad 0.2,\ S_{21} = \quad 5.0$$
$$S_{12} = -0.1,\ S_{22} = -0.5$$

(a) Find the impedance Z that will maximize the power delivered to the load, Z_L.

(b) Find the power that will be delivered to the load, assuming that the matching network has been designed to present the Z found in part 19a to the output of the 2-port. Express your result in dBm.

(c) Find the impedance Z that will maximize the power delivered to the input of the 2-port.

(d) Find the power that will be delivered to the load, assuming that the matching network has been designed to present the Z found in part 19c to the output of the 2-port. Express your result in dBm.

20. In this problem, we prove that a particular set of passive terminations that cause the negative resistance criterion for oscillation to be satisfied at one port will automatically cause the conditions for oscillation to be satisfied at the other port if the 2 port is bilateral (i.e. if $S_{12} \neq 0$ and $S_{21} \neq 0$).

    (a) Show that the negative resistance criterion for steady-state oscillation ($Z_1 + Z_2 = 0$) can be written in terms of the corresponding reflection coefficients as $\Gamma_1\Gamma_2 = 1$.

    (b) Now, suppose that Γ_L plane stability circle analysis shows that at least part of unstable region lies within the unit circle in the Γ_L plane. We choose a load reflection coefficient within the unstable region and also inside of the unit circle. Denote this load termination by Γ_{Lu}. Find the reflection coefficient of the passive source termination, Γ_{Su}, that will cause the conditions for steady state oscillation to be satisfied at the input port. Write your answer in terms of the S parameters and Γ_{Lu}.

    (c) Now, show that the conditions for oscillation are satisfied at the output port when the 2-port is terminated with Γ_{Lu} and Γ_{Su}.

21. A 2-port has S parameters ($Z_o = 50\Omega$): $S_{11} = 0.2,\ S_{12} = 0.0,\ S_{21} = 2.0,\ S_{22} = 0.8$

    (a) Suppose that the 2-port is used in a system with source impedance $Z_S = 50\Omega$ and load impedance $Z_L = 50\Omega$. The power available from the source $P_{avs} = 0$ dBm. Find the power that will be delivered to the load. Express your answer in dBm.

(b) Calculate the input impedance of the 2-port when it is used in the system described in part a.

(c) This 2 port is unconditionally stable. Find the power delivered to the load when lossless matching networks are used at the input and the output of the 2 port so that the 2 port is simultaneously matched at both ports. Express your answer in dBm.

(d) Find the power delivered to the load if 2 of these 2 ports are cascaded and the cascade is used in between the source and load specified in part a. Express your answer in dBm.

22. A 2 port has the following Z parameters (all given in Ω):

$$Z_{11} = 20 \;\; Z_{22} = 300 \;\; Z_{12} = 0.0 \;\; Z_{21} = 1000$$

Find S_{11} and S_{21} ($Z_o = 50\,\Omega$). Hint: the input impedance and voltage gain of a 2 port can be written in terms of Z parameters and the load impedance as follows:

$$Z_{IN} = Z_{11} - \frac{Z_{12}Z_{21}}{Z_L + Z_{22}}$$

$$A_v = \frac{V_2}{V_1} = \frac{Z_{21}Z_L}{Z_{11}Z_L + Z_{11}Z_{22} - Z_{12}Z_{21}}.$$

23. In a particular system it is found that the operating, transducer, and available gains of a 2-port are $G = 10\,\text{dB}$, $G_T = 8\,\text{dB}$, $G_A = 14\,\text{dB}$. The 2-port is unilateral and is unconditionally stable. The source impedance $Z_S = 100\,\Omega$ and the power available from the source is 3 dBm.

(a) Find the power delivered to the load. Express your result in Watts (NOT in dBm! I want to see that you know the relationship between power in Watts and dBm.)

(b) Find the power delivered to the 2-port by the source. Express your result in dBm.

(c) Find the power that would be delivered to the load if a single lossless matching network is used between the source and the 2-port. Express your result in dBm.

(d) Find the power that would be delivered to the load if a single lossless matching network is used between the 2-port and the load. Express your result in dBm.

(e) Find the power that would be delivered to the load if lossless matching networks are used at both the input and the output of the 2-port. Express your result in dBm.

(f) Suppose it is known that the 2-port has $S_{22} = 0$ ($Z_o = 100\,\Omega$) and calculate the power that would be delivered to the load if a cascade of 2 of these identical 2-ports is used between the source and the load. Express your result in dBm.

24. In a system with $P_{avs} = 0$ dBm, $Z_S = 100\,\Omega$ and unknown, but constant, Z_L, it is empirically determined that the operating, transducer, and available gains of a particular 2-port are $G = 12$ dB, $G_T = 8$ dB, $G_A = 14$ dB when this single 2-port is used to couple the source to the load. The 2-port is known to be unilateral, unconditionally stable, and to have $S_{22} = 0$ ($Z_o = 100\,\Omega$).

(a) Calculate the power that would be delivered to the load if a cascade of 2 of these identical 2-ports is used between the source and load described above. Express your result in dBm.

(b) Suppose that a lossless matching network is added between the load and the output of the second 2-port in the system described in part a. Determine the power delivered to the load. Express your result in dBm.

25. Many amplifiers are designed to be simultaneously conjugate-matched at both ports when used between $50\,\Omega$ source and load impedances. Suppose that you have obtained such an amplifier, and that the specifications for the amplifier state that the power gain of the unit is 20 dB in a $50\,\Omega$ system. Furthermore, the specifications state that the reverse isolation of the amplifier is 30 dB in a $50\,\Omega$ system. Reverse isolation is the power attenuation of the amplifier when the amplifier is driven at the output port and the load is connected to the input port. (Reverse isolation of 30 dB means that the reverse power gain of the amplifier is -30 dB.)

(a) What are S_{11} and S_{22} ($Z_o = 50\,\Omega$) for the amplifier?

(b) Suppose that the amplifier is used between a $50\,\Omega$ source and a $1000\,\Omega$ Ohm load. What is the transducer gain of the amplifier in this system? Express your result in dB.

(c) What is the largest input reflection coefficient magnitude, $|\Gamma_{IN}|$, that could ever be seen at the input of the amplifier, assuming that the amplifier is terminated with a passive load?

(d) Suppose that the amplifier is used between a $200\,\Omega$ source and a $200\,\Omega$ load. You do not have enough information to determine the exact transducer gain in this system, however you do have enough information to determine upper and lower limits for what the transducer gain could be. Specify the range of possible transducer gains (in dB).

26. Consider a 2-port consisting of a passive, lossless ladder network that matches a 50Ω source to the load $Z_L = 10 - j200\,\Omega$. Answer the following questions. Note carefully that you can answer all parts of this question without actually designing the matching network.

(a) Find $|S_{21}|$ ($Z_o = 50\,\Omega$) for the 2-port.

(b) Find $|S_{11}|$ ($Z_o = 50\,\Omega$) for the 2-port.

(c) Find S_{22} ($Z_o = 50\,\Omega$) for the 2-port. Find the magnitude and phase angle.

27. A unilateral amplifier is simultaneously conjugate-matched at both ports and has a transducer gain of 12 dB when used in a system with $Z_S = Z_L = 300\,\Omega$.

(a) What is the transducer gain (in dB) when this amplifier is used in a system with $Z_S = Z_L = 50\,\Omega$?

(b) What is the available gain (in dB) when this amplifier is used in a system with $Z_S = 50\,\Omega$?

28. Consider the unilateral hybrid-pi model shown in Figure 8.23.

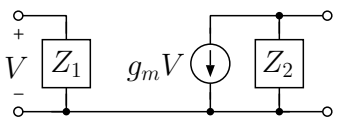


Figure 8.23:

(a) What constraints must be satisfied by Z_1 and Z_2 if the 2-port is to be unconditionally stable?

(b) Find an expression for the operating power gain of the unilateral hybrid-pi model. Express your result in terms of g_m, Z_1, Z_2, and the load impedance Z_L. It is not necessary to find S-parameters (or any other 2-port parameter set) to work this problem. Start with the definition of operating power gain.

(c) Find an expression for the maximum available gain $G_{A,max}$ for the unilateral hybrid-pi model. Express your result in terms of g_m, Z_1, and Z_2.

29. Many amplifiers are designed to be simultaneously conjugate-matched at both ports when used between $50\,\Omega$ source and load impedances. Suppose that you have obtained such an amplifier, and that the specifications for the amplifier state that the power gain of the unit is 20 dB in a $50\,\Omega$ system. Furthermore, the specifications state that the reverse isolation of the amplifier is 23 dB in a $50\,\Omega$ system. Reverse isolation is the power attenuation of the amplifier when the amplifier is driven at the output port and the load is connected to the input port. (Reverse isolation of 23 dB means that the reverse power gain of the amplifier is -23 dB.)

(a) What are S_{11} and S_{22} ($Z_o = 50\,\Omega$) for the amplifier?

(b) Is this amplifier unconditionally stable? Justify your answer.

(c) Suppose that the amplifier is used between a $400\,\Omega$ source and a $50\,\Omega$ load. The power available from the source is $P_{avs} = -10$ dBm. Find the power delivered to the load. Express your result in dBm.

(d) Suppose that the system of part c. is modified by adding a lossless impedance matching network between the *output* of the 2-port and the $50\,\Omega$ load. Find the power delivered to the load. Express your result in dBm.

(e) Suppose that two of these amplifiers are cascaded, and used between a $400\,\Omega$ source with $P_{avs} = -10$ dBm and a $50\,\Omega$ load. Find the power delivered to the load. Express your result in dBm.

Chapter 9

Filter Design

This chapter discusses the implementation of filter networks with passive, lossless components. In particular, we will concentrate on the design of filters based on ladder networks comprised of lossless inductors and capacitors. A discussion of general properties of lossless filters in terms of the scattering parameters of the filter network is followed by a description of some useful functions which approximate the ideal rectangular lowpass response and are often used as target filter response functions when designing practical filters. Then we will look at how to design lowpass filters with a prescribed frequency dependence for the transducer power gain function, $G_T(\omega)$. Finally, we will discuss two methods for transforming an existing lowpass filter design into a bandpass filter. The first method is based on a straightforward replacement of the series inductors and shunt capacitors in a lowpass filter with series and parallel resonators, respectively. The second approach involves replacing all of the lowpass filter elements with resonators of the same type which results in a *coupled-resonator* filter.

We assume that the filter network is passive, linear, and lossless, is driven with a source having real impedance Z_o, and is terminated with real load impedance Z_o, as shown in Figure 9.1. Since $\Gamma_S = \Gamma_L = 0$ in this system, the transducer, operating, and available

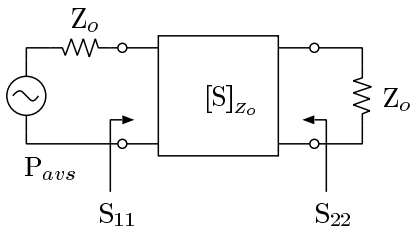


Figure 9.1: A passive, linear, lossless 2 port in a system with source and load impedances equal to Z_o.

power gains can then be written as follows:

$$G_T = \frac{P_{out}}{P_{avs}} = |S_{21}|^2 \tag{9.1}$$

$$G = \frac{P_{out}}{P_{in}} = 1 = \frac{|S_{21}|^2}{1 - |S_{11}|^2} \tag{9.2}$$

$$G_A = \frac{P_{avo}}{P_{avs}} = 1 = \frac{|S_{21}|^2}{1 - |S_{22}|^2} \tag{9.3}$$

In equations 9.2 and 9.3 we have used the fact that the filter network is lossless, which means that the operating and available power gains will be equal to one. From equations 9.2 and 9.3:

$$|S_{21}|^2 = 1 - |S_{11}|^2 = 1 - |S_{22}|^2. \tag{9.4}$$

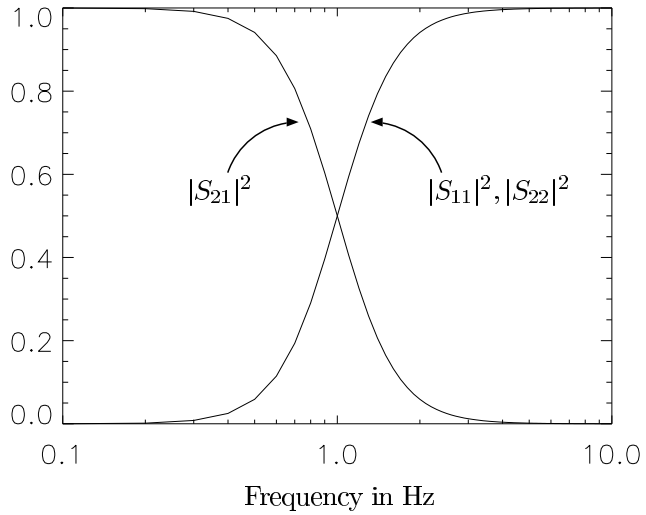


Figure 9.2: The transducer power gain, $|S_{21}|^2$, and input/output reflection coefficients for a second order Butterworth filter implemented using a passive lossless network.

Equation 9.4 shows that the transducer gain frequency response function is related to the squared magnitude of the input and output reflection coefficients of the lossless two-port. Figure 9.2 shows the transducer gain and and squared magnitude of the input coefficient reflection for a lowpass filter. Notice that the filter's attenuation at high frequencies happens because the input reflection coefficient approaches 1, i.e., the attenuation at high frequencies occurs because the source is unable to deliver much power to the network and not because of any signal absorption within the network itself. The relationship between transducer power gain and the squared magnitude of the input reflection coefficient means that the design of lossless filters can be accomplished by designing a network to realize the target transducer power gain ($|S_{21}|^2$) function or by designing the network to realize the corresponding input or output power reflection coefficient function ($|S_{11}|^2$ or $|S_{22}|^2$).

Recall that S_{11} is related to the input impedance of the network through

$$S_{11} = \frac{Z_{in} - Z_o}{Z_{in} + Z_o}|_{Z_L = Z_o}. \tag{9.5}$$

Notice that a filter with prescribed transfer gain function specified by $|S_{21}(j\omega)|^2$ can be realized by synthesizing the appropriate transducer voltage gain function, $S_{21}(j\omega)$, or equivalently, by synthesizing an input impedance function, $Z_{in}(j\omega)$, that produces $|S_{11}(j\omega)|^2 = 1 - |S_{21}(j\omega)|^2$. Both approaches will be illustrated through an example, but first we shall examine some common functions that are used for the target transducer power gain function.

9.1 Butterworth, Chebyshev, Bessel-Thompson Filters

9.1.1 Butterworth

The Butterworth response function with cutoff frequency ω_c has the following form:

$$|S_{21}(j\omega)|^2 = \frac{1}{1 + (\frac{\omega}{\omega_c})^{2n}} \tag{9.6}$$

The -3 dB frequency occurs at $\omega = \omega_c$ and the shape of the function is defined by the parameter, n, which is always an integer. The parameter n is called the order of the filter because, as we shall see, it is the highest power in the denominator polynomial of the transducer voltage transfer function, $S_{21}(j\omega)$. The Butterworth transducer gain functions for $\omega_c = 1$ and orders $n = 1$ through $n = 6$ are plotted in Figure 9.3. The tranducer gain falls off more rapidly for larger filter order n.

The Butterworth filter response approaches the ideal rectangular lowpass filter response function as the order, n, increases. The Butterworth response is called "maximally flat" because all derivatives up through order $2n - 1$ are equal to zero at $\omega = 0$ and as $\omega \to \infty$. The Butterworth approximation to the ideal rectangular response function is used when flatness of the transfer function magnitude within the passband is the highest priority for a particular application.

9.1.2 Chebyshev

The Chebyshev response function has the following form:

$$|S_{21}(j\omega)|^2 = \frac{1}{1 + \epsilon^2 C_n^2(\frac{\omega}{\omega_c})} \tag{9.7}$$

where ϵ is a constant called the "ripple" parameter, and $C_n(\omega)$ is a Chebyshev polynomial defined by:

$$\begin{aligned} C_n(\omega) = 2^{n-1}[\omega^n - \tfrac{n}{1!2^2}\omega^{n-2} + \tfrac{n(n-3)}{2!2^4}\omega^{n-4} - \tfrac{n(n-4)(n-5)}{3!2^4}\omega^{n-6} + \\ \tfrac{n(n-5)(n-6)(n-7)}{4!2^8}\omega^{n-8} - \tfrac{n(n-6)(n-7)(n-8)(n-9)}{5!2^{10}}\omega^{n-10}...] \end{aligned} \tag{9.8}$$

The summation is stopped when the exponents of ω are no longer positive. The Chebyshev polynomials for n=1 through 5 are given in Table 9.1.

Two of the Chebyshev polynomials (n=4 and n=5) are plotted in Figure 9.4.

The Chebyshev response function is plotted for $\omega_c = 1$, $\epsilon^2 = 0.0233$ and n=1 through n=4 in Figure 9.5, for $\epsilon^2 = 0.259$ and n=1 through n=4 in Figure 9.6, and for $\epsilon^2 = 0.585$ and n=1 through n=4 in Figure 9.7.

As the figures show, the ripple parameter, ϵ, controls the amount of amplitude ripple within the filter's passband. In general, for a particular filter order, n, a larger value for

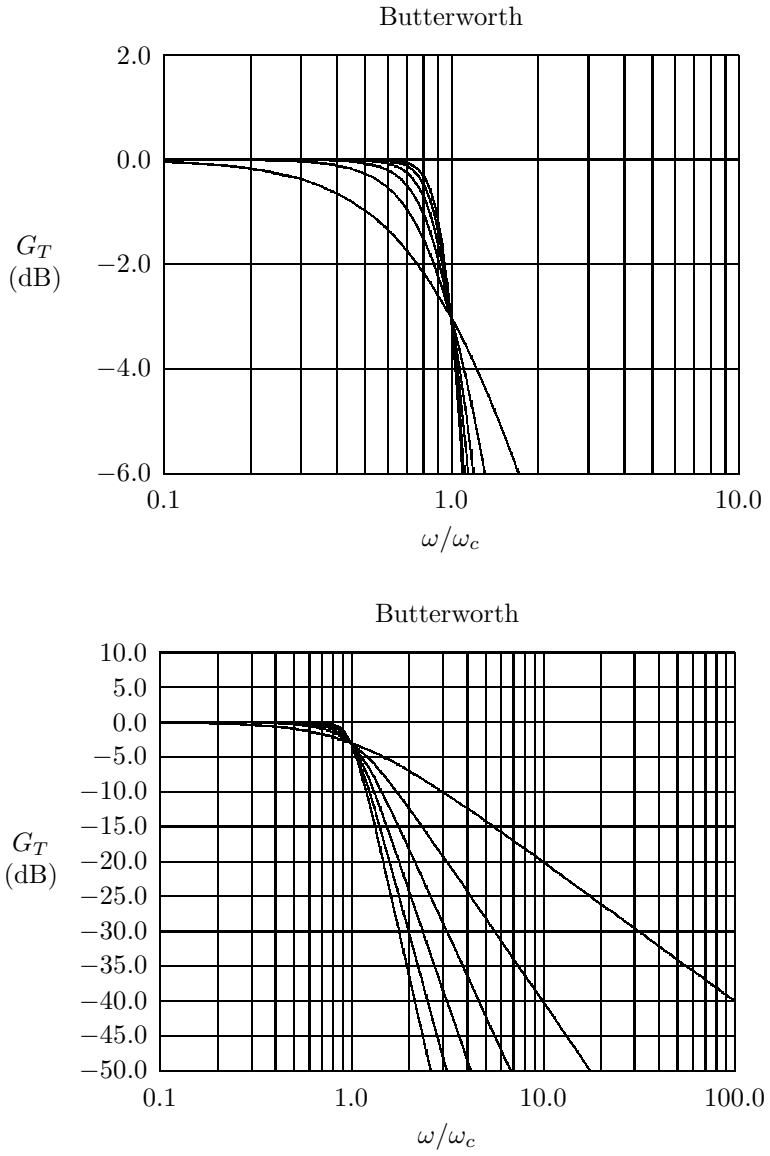


Figure 9.3: Butterworth Response function for n=1, 2, 3, 4. 5, and 6. The parameters are the same for both figures, but the lower figure shows a wider range of frequencies and gains. The higher filter orders correspond to faster descent into the stopband.

n	$C_n(x)$
1	x
2	$2x^2 - 1$
3	$4x^3 - 3x$
4	$8x^4 - 8x^2 + 1$
5	$16x^5 - 20x^3 + 5x$
6	$32x^6 - 48x^4 + 18x^2 - 1$
7	$64x^7 - 112x^5 + 56x^3 - 7x$

Table 9.1: Chebyshev polynomials for n=1 through 5.

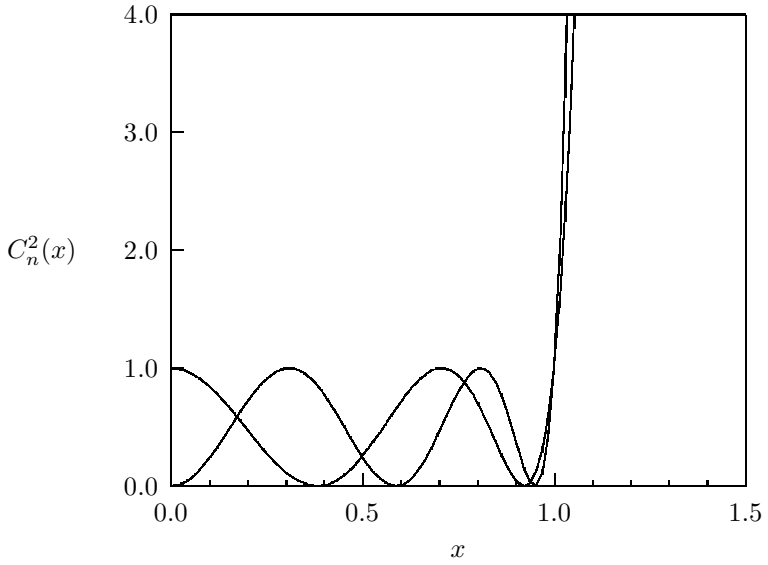


Figure 9.4: $C_n^2(x)$ for $n = 4$ and $n = 5$. Within the passband (x<1) the functions oscillate between 0 and 1. The number of extrema within the passband is equal to the filter order n. Notice that the even order polynomial ($n = 4$) is equal to one at $x = 0$. This corresponds to finite attenuation at $\omega = 0$ and will not be realizable using a lossless network with equal source and load resistances.

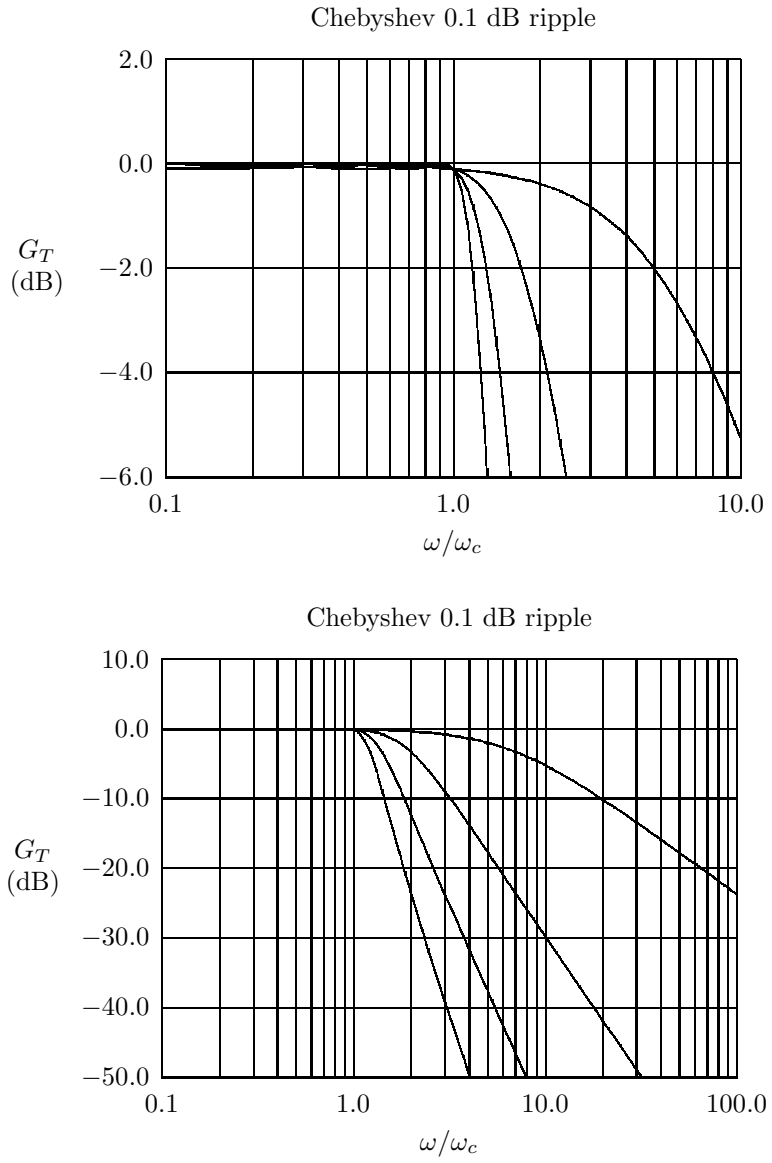


Figure 9.5: Chebyshev response; $\epsilon^2 = 0.0233$ (0.1 dB ripple), n=1, 2, 3 and 4. The lower figure shows a wider range of frequencies and gains.

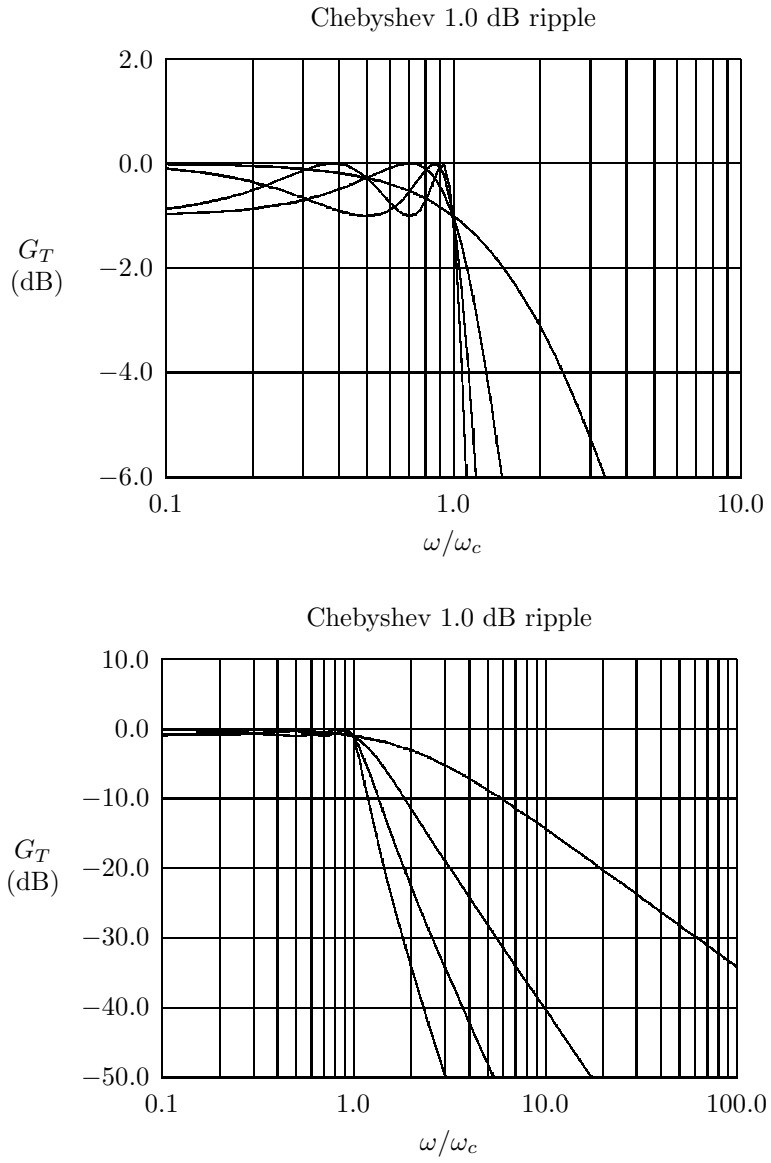


Figure 9.6: Chebyshev response; $\epsilon^2 = 0.259$ (1.0 dB ripple), n=1, 2, 3 and 4. The lower figure shows a wider range of frequencies and gains.

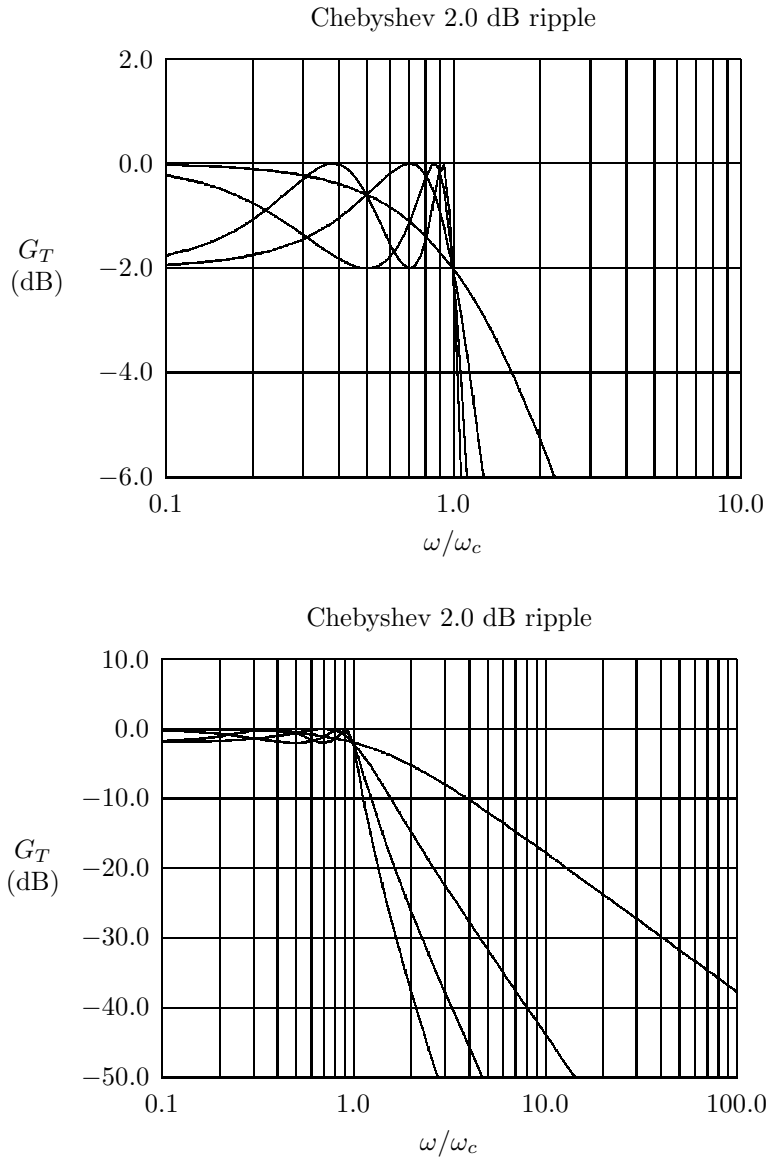


Figure 9.7: Chebyshev response; $\epsilon^2 = 0.585$ (2.0 dB ripple), n=1, 2, 3 and 4. The lower figure shows a wider range of frequencies and gains.

the ripple parameter results in a quicker roll-off into the stopband at the expense of larger ripple within the passband. The Chebyshev polynomials $C_n(\omega)$ oscillate between -1 and 1 for $|\omega| < 1$, so the transducer power gain oscillates between 1 and $\frac{1}{1+\epsilon^2}$. The ratio of the maximum and minimum responses within the passband is therefore $1 + \epsilon^2$. The ripple amplitude expressed in dB is $10 \log(1 + \epsilon^2)$. In practice, the ripple parameter is usually specified by giving the ripple amplitude in dB. Thus, a Chebyshev filter with 0.5 dB ripple corresponds to $\epsilon^2 = 10^{0.5/10} - 1$, or $\epsilon = .349$, approximately.

The cutoff frequency defined in equation 9.7 is not the -3 dB frequency. Instead, it is the frequency where the response function crosses the level corresponding to the bottom of the passband ripple on its descent into the stopband. This can be seen clearly in Figures 9.5 through 9.7.

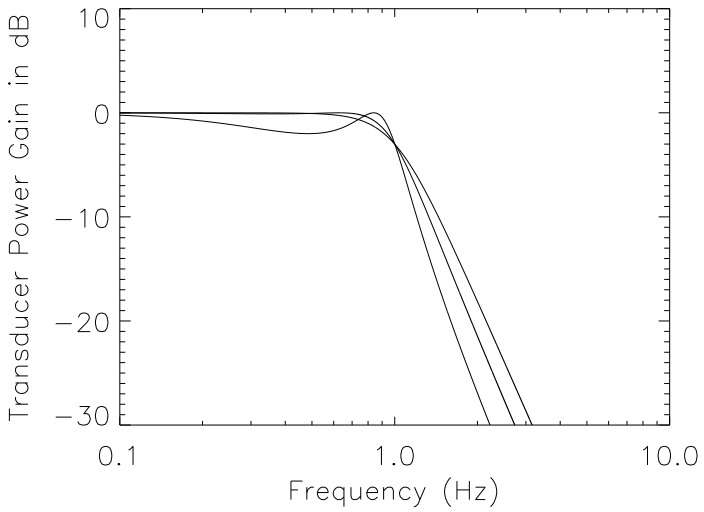


Figure 9.8: Comparison of third order Butterworth and Chebyshev filters with 0.1 dB and 2.0 dB ripple. The 2.0 dB ripple Chebyshev filter exhibits the fastest descent into the stop band. The 0.1 dB ripple Chebyshev filter has a slower descent into the stop band, but it is still noticeably faster than that of the Butterworth filter.

As the ripple parameter is decreased so that the passband ripple approaches 0 dB, the Chebyshev response approaches a maximally-flat response characteristic and becomes identical to the Butterworth response, provided that the cutoff frequency is suitably scaled. Figure 9.8 shows the n=3 response functions for Butterworth and for Chebyshev with 0.1 dB and 2.0 dB ripple, respectively. In this plot, the frequency axis has been scaled so that the -3 dB frequency of all three response functions occurs at $\omega = 1.0$. The stopband attenuation is smallest for the Butterworth response and is largest for the Chebyshev response with 2.0 dB ripple. This figure illustrates that, for a given filter order n, larger stopband attenuation can be achieved if larger ripple can be tolerated within the passband. Notice also that the 0.1 dB ripple Chebyshev filter exhibits a noticeable improvement in stopband attenuation compared to the Butterworth response, so a significant improvement in stopband attenuation can be

gained by tolerating a relatively small amount of passband ripple.

Finally, notice that the even order Chebyshev tranducer gain functions are not equal to 1 at $\omega = 0$. Instead, the even order transducer gain function is equal to $[1+\epsilon^2]^{-1}$ at $\omega = 0$, i.e. the filter has finite attenuation at $\omega = 0$. It is not possible to realize this type of transfer function with a lossless lowpass ladder network and equal source and load resistances. Since the lowpass network filter network reduces to a direct connection between the source and load terminations at $\omega = 0$, in order to have finite attenuation at $\omega = 0$ it is necessary to have different source and load terminations, such that the mismatch loss between the source and load terminations is equal to $[1+\epsilon^2]^{-1}$. If the source and load impedances are the same, as assumed here, it is only possible to realize the odd order Chebyshev response functions with a lowpass ladder network.

9.1.3 Bessel-Thompson

The Bessel-Thompson response function is desirable in some applications because it results in a group delay function that is maximally flat in the same sense that the Butterworth response provides maximally flat amplitude response. This results in nearly distortionless transmission of pulse-type waveforms. Bessel filters are commonly employed in digital communications and radar systems where it is necessary to employ a filter that will not smear pulses out over long times.

The family of Bessel transfer functions can be indexed by an integer parameter, n, and for each n the response function takes the form:

$$S_{21}(s) = \frac{B_n(0)}{B_n(s)} \tag{9.9}$$

where $B_n(x)$ are the Bessel polynomials, which satisfy

$$B_n(x) = (2n-1)B_{n-1}(x) + x^2 B_{n-2}(x)$$

with

$$B_0(x) = 1, \quad B_1(x) = x + 1.$$

Bessel polynomials for n=1 through n=5 are tabulated in Table 9.2.

n	$B_n(x)$
1	$x+1$
2	x^2+3x+3
3	$x^3+6x^2+15x+15$
4	$x^4+10x^3+45x^2+105x+105$
5	$x^5+15x^4+105x^3+420x^2+945x+945$
6	$x^6+21x^5+210x^4+1260x^3+4725x^2+10395x+10395$
7	$x^7+28x^6+378x^5+3150x^4+17325x^3+62370x^2+62370x^2+135135x+135135$

Table 9.2: Bessel polynomials for orders n=1 through n=4.

The Bessel response for order n can be written in the form:

$$S_{21}(s) = \frac{1}{1 + a_1 s + a_2 s^2 + a_3 s^3 + \cdots a_n s^n} \tag{9.10}$$

The group delay of a filter is defined as the negative slope of the filter's phase response. Thus, writing the transfer function $S_{21}(j\omega) = A(\omega)e^{j\phi(\omega)}$, the group delay, $T_g(\omega)$, is defined as:

$$T_g(\omega) = -\frac{d\phi(\omega)}{d\omega} \tag{9.11}$$

For a transfer function of the form given in equation 9.10, it can be shown that the group delay at $\omega = 0$ is equal to the value of the coefficient a_1. If the Bessel polynomials given in Table 9.2 are used in equation 9.9 and the resulting equation is manipulated so that it is in the form of equation 9.10, the coefficient a_1 will be equal to one, corresponding to a group delay of 1 second. To scale a Bessel filter for a specified delay, it is necessary to scale the coefficients a_1 through a_n by multiplying the $i'th$ coefficient by τ^i, where τ is the desired delay. Thus, a third-order Bessel filter with delay of 1 μs would have coefficients:

$$\begin{aligned} a_1 &= 10^{-6} \\ a_2 &= (10^{-6})^2 \frac{6}{15} \\ a_3 &= (10^{-6})^3 \frac{1}{15} \end{aligned}$$

The Bessel response for n=1 through n=4 is plotted in Figure 9.9. The coefficients of the Bessel polynomial have been scaled differently for each plot in order to make the -3 dB frequencies equal to 1 Hz in all cases. The Bessel response exhibits a gradual descent into the stopband and, for a given filter order and -3 dB frequency, results in less attenuation in the stopband than the Bessel or Chebyshev filters.

The group delay for the filters with n=1 through n=4 is shown in Figure 9.10. Notice that the shape of the Bessel filter's group delay curve is the same as the Butterworth filter's maximally flat gain response. The same (scaled) coefficients that were used to produce Figure 9.9 were used to produce Figure 9.10. Notice that when all of the filters are scaled to have the same -3 dB frequency, as is the case here, the delay increases as the filter order is increased.

9.2 Example: Synthesis of 4'th order Butterworth filter

Suppose that it is necessary to design a passive LC ladder network that realizes the fourth order Butterworth lowpass response function, i.e., the network transducer gain function must be

$$|S_{21}(\omega)|^2 = \frac{1}{1+\omega^8} \tag{9.12}$$

Notice that the cutoff frequency of the filter has been set to 1 rad/s. Later we will discuss how to scale the design to an arbitrary cutoff frequency.

So far, only the magnitude of the transfer function has been specified. To design an actual circuit that realizes the desired transfer function, we must determine the complex voltage transfer function, $S_{21}(j\omega)$. Alternatively, we could determine the required complex input reflection coefficient, $S_{11}(j\omega)$, from which we can solve for the complex input impedance function that the network must realize.

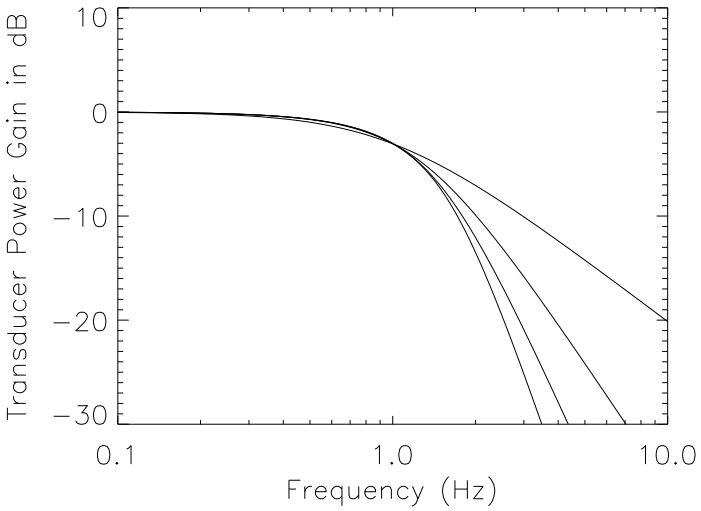


Figure 9.9: Bessel response for n=1 through n=4. The higher filter orders correspond to faster descent into the stopband.The coefficients of the Bessel polynomial given in Table 9.2 have been scaled for each curve so that the -3 dB frequency occurs at 1 Hz.

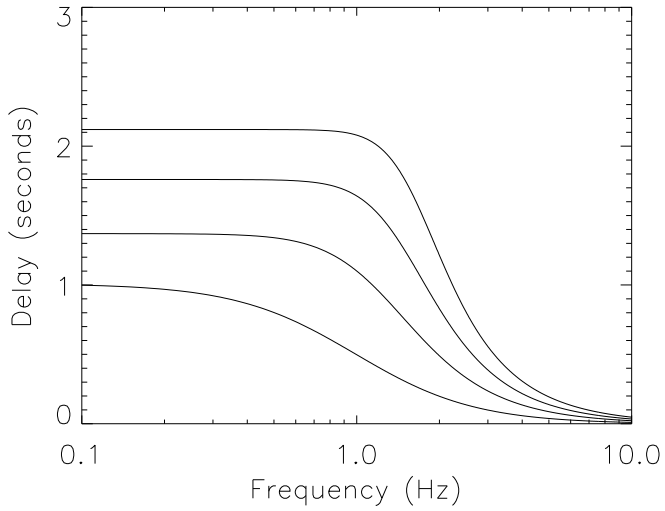


Figure 9.10: Group delay for the Bessel filters with n=1 through n=4.

9.2.1 Filter synthesis based on the S_{21} function.

We must first determine the form of a realizable S_{21} function that will provide the desired transducer gain response. The may be accomplished by noting that equation 9.12 can be re-written as follows:

$$S_{21}(j\omega)S_{21}^*(j\omega) = S_{21}(j\omega)S_{21}(-j\omega) = \frac{1}{1+\omega^8}$$

or, with $\omega = \frac{s}{j}$:

$$S_{21}(s)S_{21}(-s) = \frac{1}{1+(\frac{s}{j})^8} = \frac{1}{1+s^8} \tag{9.13}$$

The right hand side of equation 9.13 has 8 poles. To determine $S_{21}(s)$, the right hand side of equation 9.13 must be factored into a product of two terms such that the second term can be obtained by replacing s with $-s$ in the first term. This amounts to deciding which group of 4 poles must be assigned to $S_{21}(s)$. It will be necessary to ensure that the poles assigned to $S_{21}(s)$ are in the left half of the s-plane. The poles must be in the left-half plane because it is not possible to realize an S_{21} function that has poles in the right-half plane using a passive network. The pole locations are given by the roots of the following equation:

$$s^8 = -1 \tag{9.14}$$

$$s^8 = e^{j(2n+1)\pi} \tag{9.15}$$

The solutions to equation 9.14 (or, equivalently, equation 9.15) are $s_k = e^{j(2n+1)\frac{\pi}{8}}$ with $n = 0, 1, 2, \cdots 7$. Thus, the poles are equally spaced around the unit circle centered on the origin of the s-plane. The 4 poles that must be assigned to $S_{21}(s)$ are those in the left-half plane. The pole locations can be written as follows:

$$\begin{aligned} s_k &= e^{j(2n+1)\frac{\pi}{8}} \\ &= \cos(\frac{(2n+1)\pi}{8}) + j\sin(\frac{(2n+1)\pi}{8}) \end{aligned}$$

with the left-half plane pole locations obtained when $n = 2, 3, 4, 5$. $S_{21}(s)$ can now be written as follows:

$$S_{21}(s) = \frac{\pm 1}{(s-s_2)(s-s_3)(s-s_4)(s-s_5)} \tag{9.16}$$

$$S_{21}(s) = \frac{\pm 1}{(s-e^{j\frac{5\pi}{8}})(s-e^{j\frac{7\pi}{8}})(s-e^{j\frac{9\pi}{8}})(s-e^{j\frac{11\pi}{8}})}$$

The denominator can be expanded to yield:

$$S_{21}(s) = \frac{\pm 1}{s^4 + 2.61313s^3 + 3.41421s^2 + 2.61313s + 1}. \tag{9.17}$$

The ladder networks that can realize the transfer function with 4 poles will have 4 branches with one energy storage element per branch. The two possibilities are the lowpass

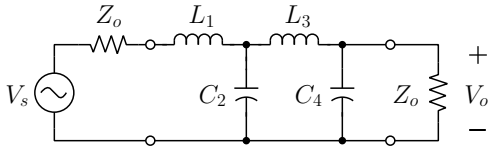


Figure 9.11:

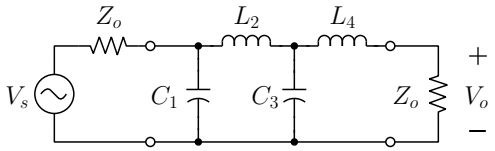


Figure 9.12:

networks shown in Figures 9.11 and 9.12. Note that in the limit as ω approaches 0, these networks will have the property that $S_{21} \to 1$. Thus, only the upper sign in the numerator of equations 9.16 and 9.17 is relevant.

Consider Figure 9.11. If the termination impedances are 1 Ω, the transducer voltage gain for this network is easily calculated. (Assume that the voltage across the load is 1 V and work backwords toward the source to find V_S. Then, $S_{21} = 2/V_S$.) The result is:

$$S_{21} = \frac{1}{s^4\left(\frac{L_1C_2L_3C_4}{2}\right) + s^3\left(\frac{L_1C_2L_3+C_2L_3C_4}{2}\right) + s^2\left(\frac{L_3C_4+L_1C_4+L_1C_2+C_2L_3}{2}\right) + s\left(\frac{L_3+L_1+C_2+C_4}{2}\right) + 1} \tag{9.18}$$

Comparing the coefficients in the denominator of 9.17 with those in the denominator of 9.18 yields four equations for the four unknown parameters:

$$\frac{L_1C_2L_3C_4}{2} = 1 \tag{9.19}$$

$$\frac{L_1C_2L_3 + C_2L_3C_4}{2} = 2.61313 \tag{9.20}$$

$$\frac{L_3C_4 + L_1C_4 + L_1C_2 + C_2L_3}{2} = 3.41421 \tag{9.21}$$

$$\frac{L_1 + L_3 + C_2 + C_4}{2} = 2.61313 \tag{9.22}$$

Equations 9.19 through 9.22 can now be solved for the unknown parameters. The solution

is:

$$\begin{aligned} L_1 &= 0.7654\,\mathrm{H} \\ C_2 &= 1.8478\,\mathrm{F} \\ L_3 &= 1.8478\,\mathrm{H} \\ C_4 &= 0.7654\,\mathrm{F} \end{aligned} \tag{9.23}$$

These values will give a Butterworth response with cutoff frequency $\omega_c = 1$ rad/s if the terminating impedances are 1 Ω. This filter is referred to as a lowpass prototype filter and can be used as the basis for a filter with arbitrary cutoff frequency and termination impedance.

9.2.2 Filter synthesis based on the input impedance function

From 9.4 we have

$$|S_{11}(j\omega)|^2 = 1 - |S_{21}(j\omega)|^2$$

or, with $s = j\omega$:

$$S_{11}(s)S_{11}(-s) = 1 - \frac{1}{1 + (\frac{s}{j})^8} = \frac{s^8}{1 + s^8}.$$

As in the previous section we must factor the RHS to isolate a term having poles only in the left-half plane, because an S_{11} function that can be realized with passive network will have poles only in the left-half plane. Proceeding exactly as before, we determine that

$$S_{11}(s) = \frac{\pm s^4}{(s - s_2)(s - s_3)(s - s_4)(s - s_5)} \tag{9.24}$$

$$S_{11}(s) = \frac{\pm s^4}{(s - e^{j\frac{5\pi}{8}})(s - e^{j\frac{7\pi}{8}})(s - e^{j\frac{9\pi}{8}})(s - e^{j\frac{11\pi}{8}})}$$

Expanding the denominator, as before:

$$S_{11}(s) = \frac{\pm s^4}{s^4 + 2.61313s^3 + 3.41421s^2 + 2.61313s + 1}. \tag{9.25}$$

The input impedance function, assuming 1 Ω source and load impedance, is

$$z_{in}(s) = \frac{1 + S_{11}(s)}{1 - S_{11}(s)}$$

Taking the upper (plus) sign in the numerator of 9.25

$$z_{in}(s) = \frac{2s^4 + 2.61313s^3 + 3.41421s^2 + 2.61313s + 1}{2.61313s^3 + 3.41421s^2 + 2.61313s + 1}.$$

Using long division, the input impedance can be written in continued fraction form

$$z_{in}(s) = 0.7654s + \frac{1}{1.8478s + \frac{1}{1.8478s + \frac{1}{.7654s + 1}}}. \tag{9.26}$$

Next, notice that the input impedance of the network shown in Figure 9.11 can be written in continued fraction form as:

$$z_{in}(s) = sL_1 + \frac{1}{sC_2 + \frac{1}{sL_3 + \frac{1}{sC_4+1}}}. \tag{9.27}$$

Equating the coefficients in equations 9.27 and 9.26 yields the results already given in equation 9.23.

If the lower (minus) sign is selected in the numerator of equation 9.25 the expression for $z_{in}(s)$ will be the inverse of that given in equation 9.26. The input admittance for that case would be equal to the right hand side of equation 9.26 and the continued fraction expansion representation of the input admittance will correspond to the right hand side of equation 9.26. The ladder network shown in Figure 9.12 has an input admittance function of the same form. Equating coefficients as before yields the normalized component values for the ladder network of the form shown in Figure 9.12:

$$\begin{aligned} C_1 &= 0.7654\,\mathrm{F} \\ L_2 &= 1.8478\,\mathrm{H} \\ C_3 &= 1.8478\,\mathrm{F} \\ L_4 &= 0.7654\,\mathrm{H} \end{aligned} \tag{9.28}$$

Notice that the same list of normalized component values applies to both of the networks shown in Figures 9.11 and 9.12, but the values of the inductances in one case correspond to the values of the capacitors in the other case, and vice versa. Therefore, when tabulating the normalized component values for filters of a given type and order it is sufficient to give one list of component values. The elements in the list are interpreted as alternating inductance and capacitance values, starting with the left side of the network, if the ladder network has a series inductor on the left side of the network (as in Figure 9.11). When applied to a ladder network with a shunt capacitor on the left side of the network (as in Figure 9.12) the elements in the list are interpreted as alternating capacitance and inductance values.

The element values for lowpass prototype filters of any type and order can be derived by applying the procedures illustrated in this example. The essential information that is required to derive the component values is the locations of the left-half plane poles associated with the Transducer power gain function. A table of component values for lowpass prototype Butterworth filters is provided in section 9.2.3.

9.2.3 Component values for lowpass prototype Butterworth filters

n	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
1	2.0									
2	1.41421	1.41421								
3	1.0	2.0	1.0							
4	0.765367	1.84776	1.84774	0.765367						
5	0.618034	1.61803	2.0	1.61803	0.618034					
6	0.517638	1.41421	1.93185	1.93185	1.41421	0.517638				
7	0.445042	1.24698	1.80194	2.0	1.80194	1.24698	0.445042			
8	0.390181	1.11114	1.66294	1.96157	1.96157	1.66294	1.11114	0.390181		
9	0.347296	1.0	1.53209	1.87939	2.0	1.87939	1.53209	1.0	0.347296	
10	0.312869	0.907981	1.41421	1.78201	1.97538	1.97538	1.78201	1.41421	0.907981	0.312869

Table 9.3: Component values for lowpass prototype Butterworth filters.

9.3 Example - 3'rd order/0.1 dB ripple Chebyshev lowpass

The transducer power gain function for a Chebyshev filter has the form

$$|S_{21}(j\omega)|^2 = \frac{1}{1+\epsilon^2 C_n^2(\omega)}.$$

Then, with $\omega = s/j$:

$$|S_{21}(s)|^2 = \frac{1}{1+\epsilon^2 C_n^2(\frac{s}{j})}.$$

$$|S_{11}(s)|^2 = 1 - |S_{21}(s)|^2 = \frac{\epsilon^2 C_n^2(\frac{s}{j})}{1+\epsilon^2 C_n^2(\frac{s}{j})}. \tag{9.29}$$

Note that the -3 dB corner frequency of the Chebyshev response function does not occur at $\omega = 1$. Instead, $\omega = 1$ corresponds to the frequency where the response has an attenuation equal to the specified ripple. The response descends into the stopband at $\omega > 1$. After the lowpass prototype filter has been designed, the component values can be scaled to yield a new prototype filter that has -3 dB frequency equal to $\omega = 1$.

A third order Chebyshev filter with 0.1 dB ripple is obtained by using$\epsilon^2 = 10^{(0.1/10)} - 1 = 0.023293$ and $C_3^2(x) = (4x^3 - 3x)^2$ in equation 9.29, which yields:

$$|S_{11}(s)|^2 = \frac{(s^3 + 0.75s)^2}{s^6 + 1.5s^4 + 0.5625s^2 - 2.68321}$$

The left-half plane poles associated with this function are located at:

$$\begin{aligned} s_1 &= -0.969406 \\ s_2 &= -0.484703 - j1.20616 \\ s_3 &= -0.484703 + j1.20616 \end{aligned}$$

Thus, we have

$$S_{11}(s) = \frac{\pm(s^3 + 0.75s)}{(s-s_1)(s-s_2)(s-s_3)} = \frac{\pm(s^3 + 0.75s)}{s^3 + 1.93881s^2 + 2.62949s + 1.63805}$$

If the upper sign is chosen, then:

$$z_{in}(s) = \frac{1+S_{11}(s)}{1-S_{11}(s)} = \frac{2s^3 + 1.93881s^2 + 3.37949s + 1.63805}{+1.93881s^2 + 1.87848s + 1.63805}.$$

Using long division:

$$z_{in}(s) = 1.0316s + \frac{1}{1.1474s + \frac{1}{1.0316s+1}}.$$

Thus, the element values for the lowpass prototype filter are:

$$g_1 = 1.0316 \quad g_2 = 1.1474 \quad g_3 = 1.0316.$$

Since the network has 3 elements, it will be either a "T" network or a "PI" network. If the T-network is employed, g_1 and g_3 represent the values of the series inductors (in Henries) and g_2 represents the value of the shunt capacitor (in Farads). If the PI-network is employed g_1 and g_3 are the shunt capacitor values (in Farads) and g_2 is the series inductor value (in Henries).

It may be desirable to scale the Chebyshev prototype filters such that the -3 dB frequency is located at $\omega = 1\, s^{-1}$. The third order prototype filter that has been designed so far has attenuation equal to 0.1 dB at $\omega = 1\, s^{-1}$. The -3 dB frequency is located at $\omega = 1.3795\, s^{-1}$. The filter can be scaled so that the -3 dB frequency is at $\omega = 1$ by multiplying the element values by 1.3795. The new prototype values are:

$$g_1 = 1.4994 \quad g_2 = 1.6678 \quad g_3 = 1.4994$$

9.3.1 Component values for odd-order lowpass prototype Chebyshev filters with 0.1 dB ripple

n	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
1	0.305241								
3	1.03156	1.1474	1.03156						
5	1.14681	1.37121	1.975	1.37121	1.14681				
7	1.18118	1.42281	2.09667	1.5734	2.09667	1.42281	1.18118		
9	1.19567	1.4426	2.13455	1.61672	2.20537	1.61672	2.13455	1.4426	1.19567

Table 9.4: Component values for odd order lowpass prototype Chebyshev filters with 0.1 dB ripple and equal source and load terminations. These prototype values produce a filter with attenuation equal to the passband ripple (-0.1 dB) at $\omega = 1\,\mathrm{s}^{-1}$.

9.4 Frequency and Impedance scaling

So far we have synthesized lowpass prototype filters that have provide some desired response function having a corner frequency of $\omega = 1\,\mathrm{s}^{-1}$ when terminated with source and load resistance of $1\,\Omega$. To scale a lowpass prototype filter to a new frequency, it is necessary to scale the component values so that they have the same reactance at the new cutoff frequency as the prototype values have at $\omega = \omega_c = 1$ rad/s. The new inductor and capacitor values can therefore be obtained from the calculated values by dividing the calculated values by the desired cutoff frequency, in rad/s.

To scale the filter to a new impedance value, it is necessary to scale the component impedances so they maintain the same magnitude relative to the terminating impedance. For example, the fourth order Butterworth lowpass protoype has $L_1 = 0.7654\,\mathrm{H}$, so the reactance of L_1 is equal to $0.7654\,\Omega$ at the cutoff frequency, or 0.7654 times the terminating impedance. Thus, for a terminating impedance of $50\,\Omega$ the inductor should have a reactance of $0.7654(50) = 38.27\,\Omega$ at ω_c. So inductor values have to be multiplied by the desired terminating impedance, Z_o. The capacitor C_2 has a reactance of $1/1.8478 = 0.5412\,\Omega$ at the cutoff frequency. To scale for a terminating impedance of $50\,\Omega$ the capacitor should have a reactance of $0.5412(50) = 27.06\,\Omega$ at the cutoff frequency. Thus, the calculated capacitor value should be divided by the new terminating impedance.

Both frequency and impedance scaling can be performed in one step as follows. Denote the new scaled element values by primed quantities and the lowpass prototype values by unprimed quantities. Then the new inductor and capacitor values are given in terms of the desired termination impedance, Z_o, and the desired cutoff frequency, ω_c, by:

$$L' = \frac{LZ_o}{\omega_c} \tag{9.30}$$

$$C' = \frac{C}{\omega_c Z_o}. \tag{9.31}$$

9.5 Bandpass Transformation

Bandpass filters can be realized by transforming a lowpass prototype filter. We'll discuss one simple transformation. The idea behind this transformation is to replace each branch of the lowpass prototype with a new branch consisting of two elements. The series inductors are replaced with series LC circuits, and the shunt capacitors are replaced with parallel LC circuits. Recall that the series elements of the lowpass filter (inductors) look like short circuits at the center of the filter response (at $\omega = 0$), and the shunt elements of the lowpass filter look like open circuits at the center of the filter response. We can transfer these characteristics to any other center frequency, ω_o, by making sure that the series and parallel branches of the new filter are resonant at ω_o. Likewise, at the edge of the original filter's passband, each branch had a certain impedance (relative to the terminating impedance). If we choose the elements of the new branches such that the impedance level is the same at the desired edges of the passband, then the shape of the bandpass filter's transfer function should look like a shifted version of the lowpass filter's transfer function.

The lowpass-to-bandpass filter transformation can be defined in terms of the center frequency and fractional bandwidth of the bandpass filter. Denoting the desired upper and lower cutoff frequencies of the bandpass filter by f_l and f_u and the center frequency by f_o,

the absolute bandwidth is:

$$BW = f_u - f_l \tag{9.32}$$

and the fractional bandwidth is:

$$bw = \frac{BW}{f_o} \tag{9.33}$$

The element values for the parallel LC shunt elements in the bandpass filter are then written in terms of the original normalized lowpass prototype filter shunt capacitor value, C_{lp} as:

$$C_{bpshunt} = \frac{C_{lp}}{bw} \tag{9.34}$$

$$L_{bpshunt} = \frac{1}{C_{bpshunt}} \tag{9.35}$$

The element values for the series LC series elements in the bandpass filter are given in terms of the original normalized lowpass prototype series inductor value, L_{lp} as:

$$L_{bpseries} = \frac{L_{lp}}{bw} \tag{9.36}$$

$$C_{bpseries} = \frac{1}{L_{bpseries}} \tag{9.37}$$

These normalized bandpass filter element values are scaled to the desired center frequency and impedance value by using equations 9.30 and 9.31 with the cutoff frequency, ω_c , replaced with the desired center frequency, ω_o.

9.5.1 Example - Bandpass filter based on 4'th order Butterworth lowpass prototype

Suppose it is necessary to design a bandpass filter with center frequency of 70 MHz, 3 dB bandwidth of 20 MHz. The filter will operate with source and load impedances of 50Ω. The fractional bandwidth is then:

$$bw = \frac{20}{70} = .2857 \tag{9.38}$$

Using the 4th order lowpass Butterworth filter prototype derived in section 9.2, and transforming that filter, the resulting bandpass filter will have the topology shown in Figure 9.13.

The element values for this filter can be obtained by scaling the the lowpass prototype values for the 4th order Butterworth filter derived in section 9.2. The results are:

L_1	=	0.305 μH	C_1	=	17.0 pF
L_2	=	17.6 nH	C_2	=	294 pF
L_3	=	0.735 μH	C_3	=	7.03 pF
L_4	=	42.4 nH	C_4	=	126 pF

The transducer power gain of this filter is shown in Figure 9.14 on linear and logarithmic scales. The lower plot shows that bandpass filters derived from lowpass prototypes using

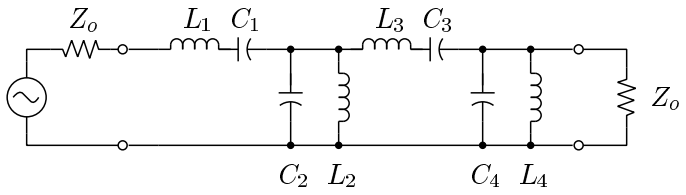


Figure 9.13: Bandpass filter topology derived from the prototype lowpass filter shown in Figure 9.11.

the transformation described in this section do not exhibit arithmetic symmetry, i.e., the response falls off more slowly above the passband than it does below the passband. The response would appear to be symmetric if plotted on a logarithmic frequency axis, as shown in Figure 9.15.

In practice, the very large stopband attenuation shown in Figure 9.15 would not be attained with a real filter. The analysis has ignored parasitic reactances associated with real components (e.g., lead inductance and capacitance between turns in an inductor). In addition, unmodeled coupling between elements and the input and output ports of the filter will inevitably occur. These effects generally limit the attainable stopband attenuation to numbers in the range 50-70 dB.

Finally, it should be noted that this method of transforming a lowpass prototype into a bandpass filter will result in realizable component values only when the target fractional bandwidth is larger than 10-15%. In general, as the target fractional bandwidth becomes smaller, the ratio between the largest and smallest required component values will increase. Notice that in this 20% bandwidth design, the ratio between the largest and smallest inductor and between the largest and smallest capacitor is 42. As the target bandwidth is decreased, the ratios increase until a point where it is no longer practical or possible to realize the needed component values with reasonably small losses.

9.6 Coupled resonator filters.

The lowpass to bandpass transformation discussed in the previous section results in a network containing multiple resonators, each of which has different L and C values. In addition, it was pointed out that the required element values span a very wide range when the fractional bandwidth of the filter is small. If a small fractional bandwidth (e.g. smaller than 10%) is needed it is generally better to use a filter topology that is based on identical coupled resonators. For example, one such topology is shown in Figure 9.16 where 4 parallel LC resonators are capacitively top-coupled to each other and to the source and load. This filter topology has the same number of resonators as the one derived in the previous section, however it provides the designer with the freedom to independently choose the properties of the resonators. This extra freedom makes it possible to create a filter in which all of the resonator inductors (or capacitors) are identical, greatly simplifying implementation of the filter.

The normalized lowpass prototype element values, g_i, provide the essential information necessary to design a bandpass filter based on coupled resonators. For detailed information

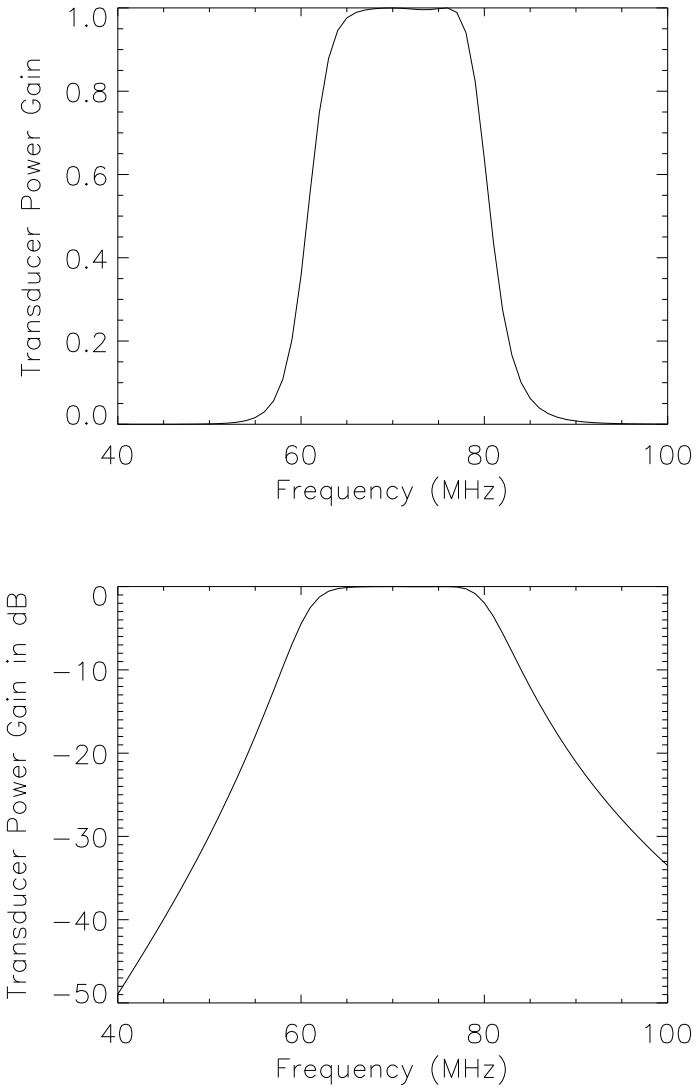


Figure 9.14: Transducer power gain ($|S_{21}|^2$) for the 70 MHz bandpass filter design based on a fourth order Butterworth lowpass prototype. The upper and lower plots show the power gain on linear, and logarithmic scales, respectively.

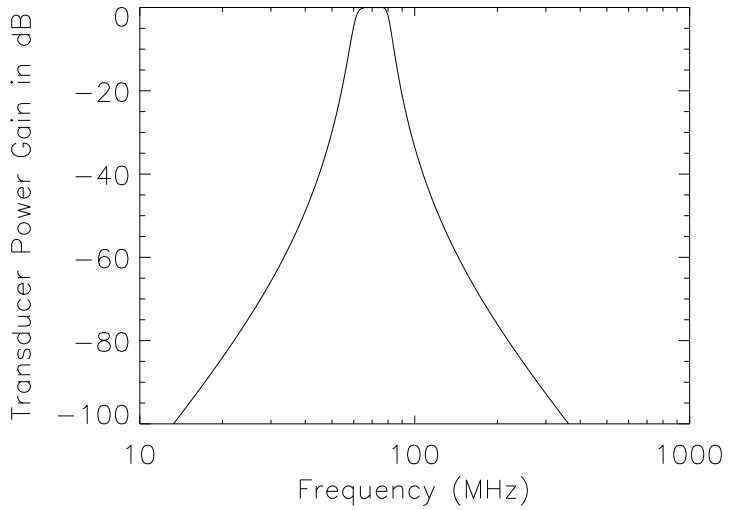


Figure 9.15: Same as Figure 9.14 except the frequency axis is plotted on a logarithmic scale to illustrate the symmetry of the response with respect to the logarithm of frequency.

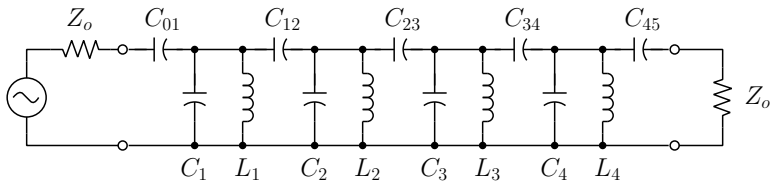


Figure 9.16: A coupled resonator filter based on capacitive coupling between parallel LC resonators.

on the design of coupled resonator filters see references [2,3,4]. Here we will simply describe the procedure for the design of a filter based on parallel LC resonators with common resonator inductance and capacitive coupling. The design proceeds as follows -

1. Determine the required corner frequencies, ω_1 and ω_2, for the prototype filter. The bandwidth of the filter will be $BW = \omega_2 - \omega_1$ and the center frequency will be $\omega_o \simeq \sqrt{\omega_1 \omega_2}$.

2. Choose a common resonator inductance, L_r, to be used for all resonators. The capacitance required to resonate this inductance at the desired center frequency, ω_o, is $C_r = (\omega_o^2 L_r)^{-1}$.

3. Calculate parameters $J_{i,i+1}$ as follows:

$$J_{0,1} = \sqrt{\frac{BWC_r}{Z_o g_1}}$$

$$J_{i,i+1}|_{i=1\,\text{to}\,n-1} = BWC_r\sqrt{\frac{1}{g_i g_{i+1}}}$$

$$J_{n,n+1} = \sqrt{\frac{BWC_r}{Z_o g_n}}$$

4. Calculate the coupling capacitor values:

$$C_{0,1} = \frac{J_{0,1}}{\omega_o\sqrt{1-(Z_o J_{0,1})^2}}$$

$$C_{i,i+1}|_{i=1\,\text{to}\,n-1} = \frac{J_{i,i+1}}{\omega_o}$$

$$C_{n,n+1} = \frac{J_{n,n+1}}{\omega_o\sqrt{1-(Z_o J_{n,n+1})^2}}.$$

5. Calculate the shunt capacitance across each resonator inductor:

$$C_1 = C_r - \frac{C_{0,1}}{1+(\omega_o Z_o C_{0,1})^2} - C_{1,2}$$

$$C_i|_{i=2\,\text{to}\,n-1} = C_r - C_{i-1,i} - C_{i,i+1}$$

$$C_n = C_r - \frac{C_{n,n+1}}{1+(\omega_o Z_o C_{n,n+1})^2} - C_{n-1,n}$$

9.6.1 Example - Coupled resonator filter with 2 parallel LC resonators based on Butterworth lowpass prototype

In this example we summarize the design of a bandpass filter based on a second order ($n = 2$) Butterworth lowpass prototype. The filter is to have center frequency 50 MHz and bandwidth 2.5 MHz (5% fractional bandwidth). It is to be used in a system with source and load impedance $Z_o = 50\,\Omega$.

To obtain a passband response that is approximately symmetric around 50 MHz, we choose the upper and lower corner frequencies as follows:

$$\omega_1 = 2\pi(50 - 2.5/2) \times 10^6 = 3.0631 \times 10^8\,\mathrm{s}^{-1}$$

$$\omega_2 = 2\pi(50 + 2.5/2) \times 10^6 = 3.2201 \times 10^8\,\mathrm{s}^{-1}.$$

Then the target center frequency will be the geometric mean of the two corner frequencies:

$$\omega_o = \sqrt{\omega_1\omega_2} = 3.1406 \times 10^8\,\mathrm{s}^{-1}$$

The target bandwidth is

$$BW = 2\pi \times 2.5 \times 10^6 = 1.5708 \times 10^7\,\mathrm{s}^{-1}.$$

The filter will contain two parallel resonators as shown in Figure 9.17. We shall choose

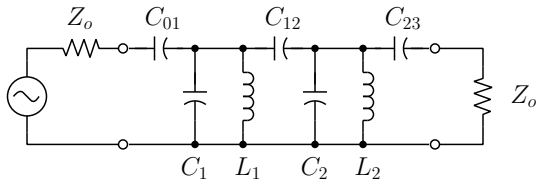


Figure 9.17: Filter topology for the coupled resonator filter design example. The design procedure that we have described assumes that the resonator inductances are equal ($L_1 = L_2 = L_r$).

a resonator inductance $L_r = L_1 = L_2 = 0.2\,\mu\mathrm{H}$. In practice, the inductance value is typically chosen by selecting the value that results in the highest possible inductor Q_L (and hence highest possible resonator Q) at the intended center frequency, ω_o. Since we ignored component losses in the design, a real filter's performance will approximate the ideal predicted performance only when the component Q of each filter component is significantly larger than the filter's $Q = f_o/BW$. The freedom to independently choose the inductance value to optimize resonator Q makes the coupled resonator filter very attractive for filters with narrow bandwidth (high Q).

The total capacitance required to resonate the chosen resonator inductance is $C_r = 1/(\omega_o^2 L_r) = 50.69\,\mathrm{pF}$. The lowpass prototype component values for a second order Butterworth filter ($n = 2$) are $g_1 = g_2 = \sqrt{2}$.

The $J_{i,i+1}$ can now be calculated. The values are:

$$J_{01} = 0.0033557 \quad J_{12} = 0.0005631 \quad J_{23} = 0.0033557.$$

The coupling capacitances are:

$$C_{01} = 10.84\,\text{pF} \quad C_{12} = 1.793\,\text{pF} \quad \text{C}_{23} = 10.84\,\text{pF}.$$

The resonator capacitances are:

$$C_1 = C_2 = 38.06\,\text{pF}.$$

The transducer power gain of the resulting filter is plotted in Figure 9.18 on both linear and logarithmic scales. The asymmetry that is evident in the bottom plot is a result of the capacitive top-coupling. At high frequencies the network reduces to an all-capacitor network with roughly constant insertion loss.

9.7 References

1. Carson, Ralph S., *Radio Concepts: Analog*, John-Wiley and Sons, New York, 1990.

2. Zverev, Anatol I., *Handbook of Filter Synthesis*, John-Wiley and Sons, New York, 1967.

3. Rhea, Randall W., *HF Filter Design and Computer Simulation*, Noble Publishing, Atlanta, 1994.

4. Matthaei, George L., Leo Young, E. M. T. Jones, *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, Artech House, Massachusetts, 1980.

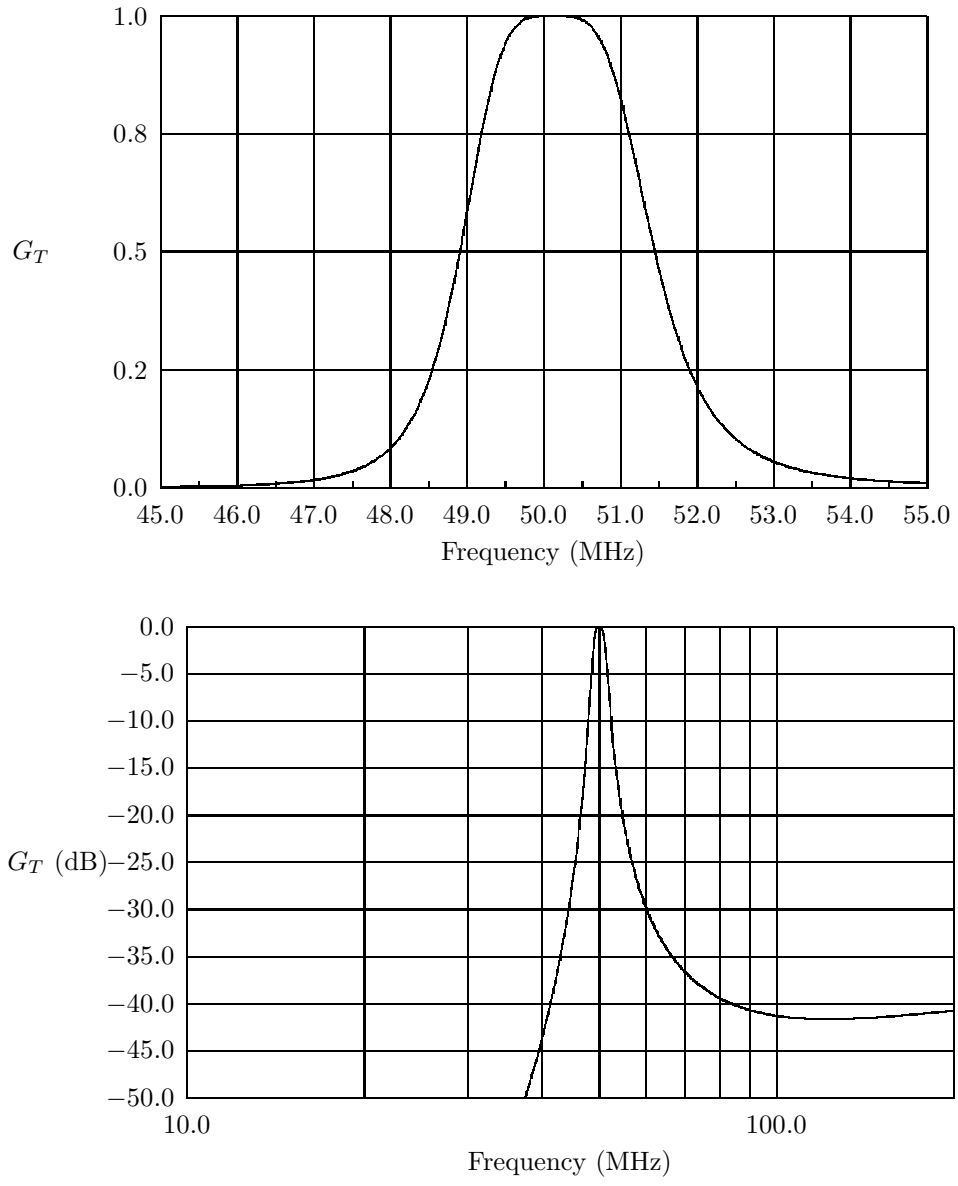


Figure 9.18: Transducer power gain ($|S_{21}|^2$) for the 50 MHz bandpass filter based on a second order Butterworth lowpass prototype and implemented with top-coupled parallel LC resonators. The upper and lower plots show the power gain on linear and logarithmic scales, respectively.

9.8 Homework Problems

1. A figure of merit that is used to characterize filters is the "shape factor," defined to be the ratio of the transducer power gain bandwidth at the -60 dB and -6 dB points, respectively. The shape factor will be larger than 1, and will approach 1 as the filter response function approaches an ideal rectangular response. Find an expression for the shape factor for Butterworth bandpass filters as a function of the filter order, n. Give numerical values for the shape factor for n=3 and n=6.

2. Find and sketch the left-half plane pole locations for a fifth-order Butterworth filter.

3. Design a second-order Butterworth lowpass filter with -3 dB frequency, 50 MHz and terminating impedances $Z_0 = 50\ \Omega$. Draw the filter and give component values. Use ADS to plot the transducer power gain ($|S_{21}|^2$). Plot the transducer power gain in dB against frequency (Use a logarithmic axis for frequency.) from at least 0.5 MHz to 500 MHz.

4. Transform the filter designed in 9-3 to a bandpass filter with center frequency 50 MHz and 3 dB bandwidth 10 MHz. Draw the filter and give component values. Use ADS to plot the transducer power gain. Plot the transducer power gain in dB against frequency on a logarithmic frequency axis from 0.5 MHz to 500 MHz. Make another plot covering a smaller frequency range around 50 MHz to show the detailed characteristics of the transfer function within the passband.

5. A Butterworth lowpass filter using the smallest possible number of components (i.e., smallest possible filter order n) is to be designed to filter the output of a source that produces output at a fundamental frequency $f_o = 15$ MHz and at harmonics of that frequency mf_0 (m = 2,3...). The attenuation at f_o should be no more than 0.5 dB and the attenuation at $2f_o$ should be at least 30 dB. Specify the -3 dB cutoff frequency of the filter f_c (in MHz) and the filter order n (an integer) that will achieve these design goals. If your parameters are used, what would the attenuation be at the second harmonic frequency $2f_o$?

Chapter 10

Noise in 1- and 2-ports

10.1 Noise Characterization of 1-ports

There are various physical phenomena that can produce "noise" in electronic circuits, including:

- *Thermal noise* - a result of random motion of charge carriers in a conductor. This type of noise will be present in any dissipative electronic circuit element (e.g. resistors or the dissipative parts of transistors and diodes), even in the absence of any externally applied bias voltage or current. Thermal noise has a "white" spectrum, which means that the power available in a frequency bandwidth *df* is independent of the center frequency of the band.
- *Shot noise* - is associated with current flow in semiconductor devices and vacuum tubes. It will be present in these active devices when an average (DC) current flows. In such devices, current flow can be modeled as a series of independent current "pulses" occurring at random and a "DC" bias current will be associated with a time-varying shot-noise current whose intensity spectrum is proportional to the DC current. The shot noise spectrum is white at sufficiently low frequencies.
- *Flicker noise or "1/f noise"* — any process with a noise power spectral density that varies as 1/f for low frequencies. If present, flicker noise will always dominate at sufficiently low frequencies.

10.1.1 Thermal Noise in Resistors

A noise voltage, as shown in Figure 10.1, is always present at the terminals of a resistor because the charge carriers within the device undergo *Brownian motion*. The random-walk of individual charges constitutes a random current, which produces a corresponding *thermal noise voltage* across the terminals of the device. This noise was identified and carefully measured by J. B. Johnson while working at Bell Telephone Laboratories[1]. A model for the noise measured by Johnson was published at the same time by H. Nyquist, who worked out an expression for the rms thermal noise voltage across any conductor. The result was

[1] "Thermal Agitation of Electricity in Conductors," J. B. Johnson, Physical Review, Vol. 32, p. 97, July, 1928.

derived using basic thermodynamics, and a fundamental result from statistical mechanics, and is known as *Nyquist's theorem.*[2]

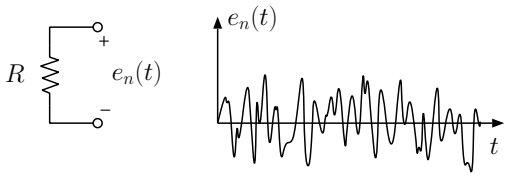


Figure 10.1: Noise voltage across the terminals of a resistor as it would be displayed on an instrument having a lowpass frequency response with bandwidth, B. The rms noise voltage is proportional to $\sqrt{B}$ and the typical time between zero crossings is proportional to B^{-1}.

The noise voltage waveform is properly modeled as a zero-mean, wide-sense-stationary, random process with Gaussian probability density function. The spectral density of the noise voltage due to thermal fluctuations is given by the *Nyquist noise formula*:

$$S_n(f) = 4R\frac{hf}{e^{hf/kT}-1} \ (\text{Volts}^2/\text{Hz}) \tag{10.1}$$

where

$$\begin{array}{rcl} f & = & \text{frequency, Hz} \\ h & = & 6.62\times10^{-34}\,\text{J}-\text{sec, Planck}'\text{s constant} \\ k & = & 1.38\times10^{-23}\,\text{J/K, Boltzman}'\text{s constant} \\ R & = & \text{resistance, } \Omega \\ T & = & \text{physical temperature of the resistor, K} \end{array} \tag{10.2}$$

When $f \ll kT/h$ so that the exponential term in the denominator of equation 10.1 is well approximated by the first two terms in a Taylor series expansion, Nyquist's formula reduces to:

$$S_n(f) = 4kTR \ (\text{Volts}^2/\text{Hz}) \tag{10.3}$$

At standard temperature, $T = 290$ K, $kT/h \simeq 6\times10^{12}$ Hz $= 6000$ GHz, which lies in the infrared. So, at standard temperature the low-frequency approximation will be valid for all RF and microwave applications. The spectral densities given in equations 10.1 and 10.3 represent the mean-square noise across the resistor within a bandwidth of 1 Hz, centered on any frequency f. In practice, the insrument used to sense the noise voltage will have a finite bandwidth that differs from 1 Hz. Suppose the instrument responds only to frequencies between the upper and lower frequencies denoted by f_l and f_h, respectively. Then the mean-square noise voltage measured by the instrument would be:

$$< e_n^2 >= \int_{f_l}^{f_h} S_n(f)df \ , \tag{10.4}$$

or, denoting the bandwidth by $B = f_h - f_l$:

$$< e_n^2 >= 4kTBR \quad (\text{Volts}^2) \tag{10.5}$$

[2]"Thermal Agitation of Electric Charge in Conductors," H. Nyquist, Physical Review, Vol. 32, p. 110, July 1928.

For example, suppose we wish to calculate the mean-square noise voltage produced across a 100 kΩ resistor in a bandwidth of 1 MHz at standard temperature (T = 290 K). At standard temperature:

$$\begin{aligned} 4kT &= 1.6\,(10^{-20})\,\text{Joules} \\ < e_n^2 > &= 4kTBR \\ &= 1.6\,(10^{-20})(10^6)(100)(10^3) \\ &= 1.6\,(10^{-9}) \quad (\text{Volts})^2 \end{aligned} \tag{10.6}$$

Thus the rms noise voltage is

$$e_{rms} = \sqrt{< e_n^2 >} = 40\,(10^{-6})\,\text{V} = 40\,\mu\text{V}. \tag{10.7}$$

The rms noise voltage across the resistor scales as $\sqrt{B}$, so the noise voltage across the resistor in a 100 MHz bandwidth would be ten times as large, or 0.4 mV.

Equation 10.5 can be generalized to describe the open circuit noise voltage across a complex impedance Z that is realized by lossless elements and resistances all having the same physical temperature. Suppose the impedance is $Z(f)$:

$$Z(f) = R(f) + jX(f), \tag{10.8}$$

then the noise variance across that impedance within bandwidth B will be

$$< e_n^2 >= 4kT \int_B R(f)\,df. \tag{10.9}$$

For example, we can calculate the total mean-square noise voltage across the parallel RC circuit shown in Figure 10.2. Here, we will make use of equation 10.9, which provides

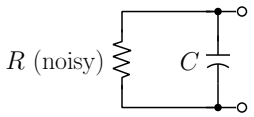


Figure 10.2: Parallel RC circuit with a noisy resistor.

a prescription for calculating the noise voltage between two terminals associated with a network consisting of an arbitrary configuration of dissipative and lossless elements. The only requirement is that all dissipative elements are at the same physical temperature. The impedance of the parallel RC combination is:

$$Z(\omega) = \frac{R(1 - j\omega RC)}{1 + (\omega RC)^2} \tag{10.10}$$

According to equation 10.9 the mean-square noise voltage depends only on the real part of the impedance:

$$\Re[Z] = R(f) = \frac{R}{1 + (2\pi f RC)^2}.$$

Using equation 10.9, the total mean-square noise voltage across the resistor is:

$$< e_n^2 >= 4kT \int_0^\infty \frac{R}{1 + (2\pi f RC)^2} df = \frac{kT}{C}.$$

Note that the result depends only on the capacitance, C, and is independent of the resistance! How is this reconciled with the fact that the mean-square voltage within any differential bandwidth, df, is proportional to the resistance? The answer is that the bandwidth of the RC filter is inversely proportional to R; so for fixed capacitance, larger resistances have larger mean-squre voltages but proportionally smaller bandwidths, hence the total noise power, integrated over all frequencies, is independent of the resistance.

A Thevenin equivalent circuit can be used to model a noisy resistor, as shown in Figure 10.3a. The Norton equivalent circuit shown in Figure 10.3b is an equivalent alternative

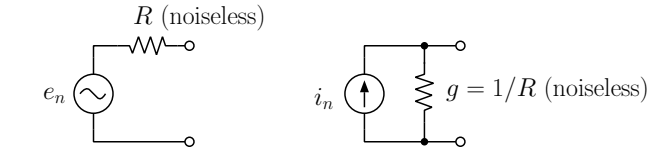


Figure 10.3: (a) Thevenin equivalent circuit. (b) Norton equivalent circuit.

representation.

The *available noise power* from a noisy resistor is defined to be the power delivered to a matched load, as in Figure 10.4. The instantaneous power delivered to the load is $\frac{(e_n/2)^2}{R}$,

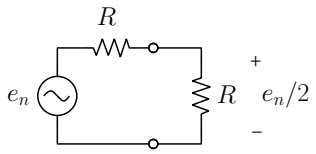


Figure 10.4: Circuit for calculation of noise power available from a noisy resistor.

so the average available noise power is $\frac{<e_n^2>}{4R}$. Using Equation 10.5, the average available noise power within a bandwidth B is:

$$\begin{aligned} \frac{< e_n^2 >}{4R} &= \frac{4kTBR}{4R} \\ &= kTB \text{ (W)} \end{aligned} \tag{10.11}$$

This result says that the noise power available from a resistor (or network of resistors all at the same physical temperature) is independent of the resistance. For later use, we will also

define the available noise power per unit bandwidth:

$$N_a = kT \text{ (W/Hz)} \tag{10.12}$$

It is useful to know that $kT_o = 4 \times 10^{-21}$ Watts, or approximately -174 dBm. So the noise power available from any resistor at standard temperature is -174 dBm per Hz.

Next, the resistor noise model is generalized to allow it to represent arbitrary one-ports, such as an antenna.

10.1.2 Noise Representation of Arbitrary 1-ports

If a 1-port device contains sources of noise in addition to the thermal noise it is said to exhibit *excess noise*. In this case it is still possible to model the device by a Thevenin or Norton equivalent circuit, and to use the Nyquist noise equation to calculate mean-square noise voltage, but it is necessary to define an effective temperature (different than the physical temperature) to account for the extra noise. When this model is used, the formula for the voltage is

$$< e_n^2 >= 4kT_nBR \tag{10.13}$$

and the noise current is

$$< i_n^2 >= 4kT_nBg \tag{10.14}$$

where R and g=1/R represent the actual resistance and conductance, but the temperature T_n is called the equivalent noise temperature of the device.

10.1.3 Noise Representation of a Receiving Antenna

Antennas are 1-ports (like resistors) and the noise power available from an antenna can be represented by an equivalent noise temperature. The real part of the antenna impedance can be written as the sum of radiation and ohmic resistances

$$\text{Re}[Z_{ant}] = R_{rad} + R_{loss} \tag{10.15}$$

where R_{rad} accounts for power radiated and R_{loss} accounts for power that is dissipated in the antenna and its nearby environment. An antenna that transmits most of the power delivered to it when used for transmission is called an efficient antenna. An efficient antenna has:

$$R_{loss} \ll R_{rad} \tag{10.16}$$

The noise voltage across an antenna's terminals comes from two sources:

1. Noise received from external sources.
2. Thermal noise generated in R_{loss}.

The noise voltage across the antenna terminals can be represented as if it was generated by an actual resistor, $R = R_{rad} + R_{loss}$, at some temperature T_A which is called the *effective antenna temperature* or *equivalent noise temperature* of the antenna. This is essentially the same as T_n, which was defined in section 10.1.2; but the subscript A is used to make it clear that the effective temperature is associated with an antenna. For an efficient antenna, T_A primarily represents noise received from external sources and has nothing to do with the physical temperature of the antenna. Instead, T_A will depend on the intensity of external

noise signals picked up by the antenna at the particular frequency of interest. In general, T_A is a parameter that must be measured, or known from previous experience with the particular antenna and location at which the antenna is to be used.

For example, suppose that an efficient antenna with radiation resistance of 200 Ω has an rms noise voltage of 0.1 μV at its terminals when measured in a bandwidth of 10^4 Hz. The equivalent noise temperature of the antenna can be calculated using:

$$e_{rms} = \sqrt{< e_n^2 >} = 0.1 \ \mu\text{V} = \sqrt{4kT_ABR}.$$

With $R = 200$ Ω and $B = 10^4$:

$$T_A = \frac{(0.1 \ \mu\text{V})^2}{4kBR} = 90.6 \text{ K}$$

Figure 10.5 shows the equivalent antenna circuit (for noise analysis).

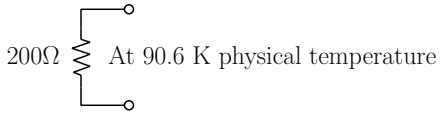


Figure 10.5: Equivalent circuit for analysis of antenna noise.

Typical antenna noise temperatures are shown in Figure 10.6. Noise radiated by sources within the earth-ionosphere cavity is the dominant source of noise below approximately 20 MHz, and is largest at night (corresponding to the curve labeled "Max.") when low ionospheric absorption allows noise to be propagated in from long distances. During the daytime, ionospheric absorption limits the amount of noise received from distant sources and the antenna temperature is significantly lower than at nighttime. Above 20 Mhz, the figure assumes that the antenna beam is directed away from the earth, so that dominant noise sources are extraterrestrial noise sources.

If we ignore radio noise generated by human activities, the primary source of natural noise between 20 MHz (roughly) and 1 GHz is radio noise emitted by sources in our galaxy and, depending on the antenna pattern and the pointing direction of the antenna relative to the sun, perhaps some noise from the sun. Antenna temperatures for antennas that look at the sky vary with time of day because the plane that contains most of the radio sources in our galaxy will drift through the antenna beam as the earth rotates. Antenna temperature due to the galactic sources also varies approximately as f^{-2}, so that maximum daily antenna temperatures might range from several thousand K at 100 MHz to a few tens of K at 1 GHz. At frequencies between 1-5 GHz or so, T_A reaches its smallest values. In fact, the contribution from sources within our galaxy falls to negligible values in this frequency range, and the dominant noise is contributed by a background noise that is uniform in all directions and corresponds to an equivalent blackbody temperature (and hence, antenna temperature) of approximately 3 K. This is the so-called *microwave background radiation* which was discovered by Bell Laboratories scientists Arno Penzias and Robert Wilson in 1965. This discovery confirmed theoretical predictions based on the "Big Bang" model of an expanding universe. Penzias and Wilson were awarded the Nobel Prize in physics in 1978, in recognition of the significance of their discovery. Their original paper had the unassuming title "A Measurement of Excess Antenna Temperature at 4080 Mc/s".

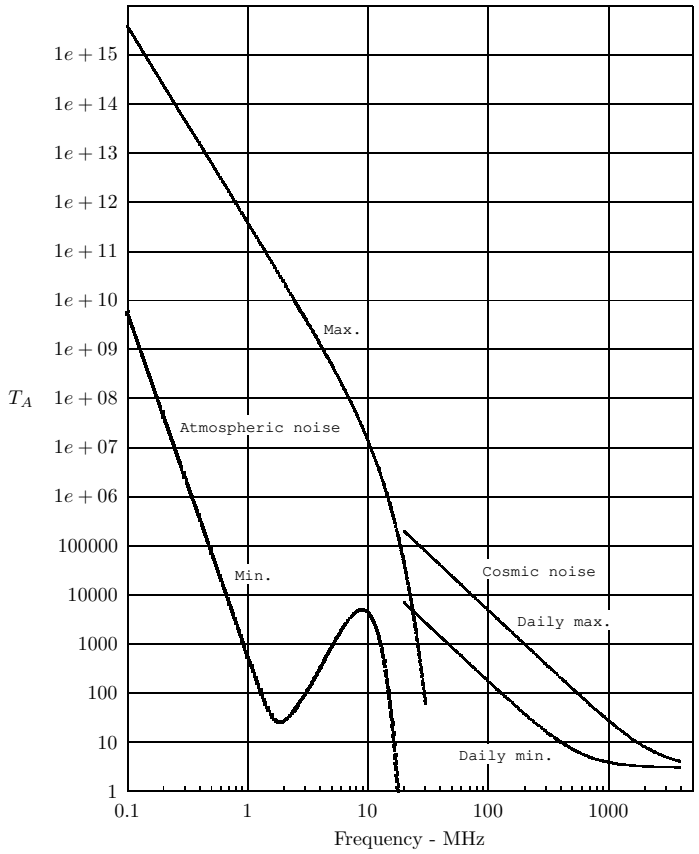


Figure 10.6: Typical antenna temperatures as a function of frequency. These curves are adapted from Figure 11-44 in the book *Electromagnetic Waves and Radiating Systems*, by Jordan and Balmain, Prentice Hall, 1968.

At frequencies above 5 GHz, antenna temperatures begin to rise again and can approach several hundred K at certain frequencies, due to thermal emission from water vapor and oxygen (O_2) in earth's atmosphere.

It can be shown that when an antenna beam looks at an object that emits thermal (black body) radiation, the antenna temperature will be equal to the temperature of the object, provided that the antenna beam is filled by the object. If the antenna beam views many objects, each having different temperatures, then the resulting antenna temperature will be a weighted average of the individual object temperatures, where the weighting depends on the angular size of each object, and the angular response function of the antenna.

When an antenna beam is pointed at the earth, as in Figure 10.7a, the antenna temperature will be approximately equal to the equivalent black body temperature of the earth or, roughly, 290 K. This has implications for satellite-to-ground communications systems. For example, when a ground station with a narrow antenna beam looks at a geostationary satellite (as shown in Figure 10.7b), the earth station antenna beam sees mostly the sky.

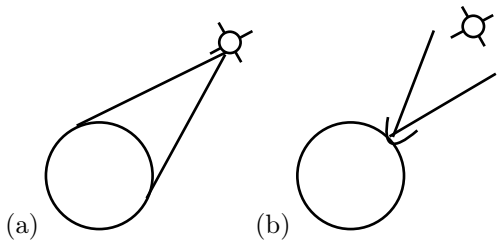


Figure 10.7: (a) When a satellite antenna beam "looks" at the earth, the beam sees a warm body, emitting thermal radiation with an effective temperature on the order of 290 K. (b) When a ground station antenna "looks" at a satellite, the beam sees mainly "cold" space.

Hence, a ground station antenna operated at a frequency between roughly 1-5 GHz will have a significantly smaller antenna temperature than a satellite antenna looking back at earth. Since the ground station antenna looks at the "cold" sky, it picks up very little external noise and the noise generated within the ground station receiver is usually a limiting factor in determining the receiver sensitivity. On the other hand, the satellite antenna looks at warm earth, and is relatively noisy and commonly contributes more noise than the receiver itself. This means that it is worthwhile to use an extremely low-noise receiver for the ground station, but not for the satellite.

For a point-to-point terrestrial communications link with the antenna radiation pattern directed horizontally over the surface of the earth, half of the beam will see cold sky, and the other half will see earth's surface, resulting in an antenna temperature somewhere in between the sky temperature and 290 K. The antenna beam may also see non-thermal sources of noise located on earth's surface, which raise the effective temperature of the earth above 290 K.

10.1.3.1 The effective antenna temperature of an inefficient antenna

Consider an antenna with feed-point resistance

$$\mathrm{Re}[Z_{ant}] = R_{rad} + R_{loss}, \tag{10.17}$$

where R_{rad} , R_{loss} represent radiation and loss components of the feedpoint resistance, respectively. *It is assumed that the presence or absence of loss does not affect the current distribution that is responsible for the radiated fields.* Hence if the antenna and all components in the vicinity of the antenna are changed to lossless materials (i.e. lossy conductors replaced with perfect conductors, lossy dielectrics replaced with perfect dielectrics) then $R_{loss} \to 0$, and R_{rad} is unaffected by this change.

Denote the antenna temperature of the (hypothetical) lossless antenna by T_A. The noise power per unit bandwidth (p.u.b.) available from the lossless antenna would be kT_A. For simplicity, assume that all lossy materials are at physical temperature T_{loss}. The noise power p.u.b. available from the loss resistance only will be kT_{loss}.

An equivalent circuit for the lossy antenna is obtained by replacing noisy resistances R_{rad} and R_{loss} with Thevenin noise voltage sources having rms voltages $\sqrt{4kT_AR_{rad}}$ and $\sqrt{4kT_{loss}R_{loss}}$ in series with noiseless resistances R_{rad} and R_{loss}, respectively. The open circuit noise voltage across the antenna terminals will have rms value $\sqrt{4k(T_AR_{rad} + T_{loss}R_{loss})}$, and the noise power available p.u.b will be

$$N_a = \frac{k(R_{rad}T_A + R_{loss}T_{loss})}{R_{rad} + R_{loss}} = k(\eta T_A + (1-\eta)T_{loss}) \tag{10.18}$$

where

$$\eta = \frac{R_{rad}}{R_{rad} + R_{loss}} \tag{10.19}$$

is the antenna efficiency. Therefore, the effective antenna temperature is

$$T_{eff} = \eta T_A + (1-\eta)T_{loss}. \tag{10.20}$$

Equation 10.20 says that as the antenna efficiency decreases ($\eta \to 0$) the effective antenna temperature aproaches the temperature of the lossy components of the antenna.

Now consider the effect of antenna losses on signal to noise ratio at the antenna output. In the lossy case, the available signal power is reduced by the factor η relative to a lossless antenna. The noise power p.u.b. available from a lossless antenna is kT_A, whereas in the lossy case it is kT_{eff} which can be written as $kT_A\eta[1 + ((1-\eta)/\eta)T_{loss}/T_A]$. Therefore the presence of loss resistance degrades the signal-to-noise power ratio at the antenna output terminals by the factor

$$\frac{SNR_{lossy}}{SNR_{lossless}} = \frac{1}{1 + \frac{1-\eta}{\eta}\frac{T_{loss}}{T_A}}. \tag{10.21}$$

This is an interesting result, because it shows that SNR degradation due to antenna losses is determined by antenna efficiency as well as the ratio T_{loss}/T_A. When $T_{loss}/T_A \gg 1$, then even a small decrease in antenna efficiency can significantly degrade the SNR. Consider a satellite ground station where T_A may be only $30\,K$. With $T_{loss} = 290\,K$, and antenna efficiency $\eta = 0.9$ (90%), the SNR is degraded by the factor 0.48, or more than 3 dB. On the other hand, if $T_{loss}/T_A \ll 1$, then even an inefficient antenna can provide essentially the same SNR as a lossless antenna. For example, the small antennas used for AM broadcast radios may have efficiencies of 1% ($\eta = 0.01$) or less, so that the factor $(1-\eta)/\eta \simeq 100$ or so. The antenna temperature, T_A is typically very large at AM broadcast frequencies, so that $T_{loss}/T_A \ll 1$. In this situation, the product $[(1-\eta)/\eta][T_{loss}/T_A]$ can be $\ll 1$ even when η itself is small, in which case the SNR available from a small inefficient antenna is essentially the same as that available from a much larger antenna with high efficiency (provided that the radiation patterns of the small and large antennas are similar).

10.2 Noise Characterization of Linear 2-ports

10.2.1 Effective Input Noise Temperature

The noise added by a 2-port (e.g., amplifier, filter, mixer) can be characterized by an effective input temperature for a given source impedance at a given frequency.

Consider in Figure 10.8 a 2-port with a hypothetical noiseless input termination, Z_s, where N_{avo} = available output noise power (the noise power delivered to a matched load).

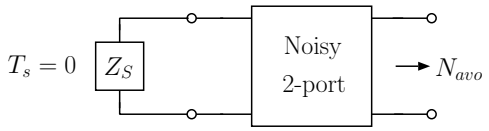


Figure 10.8: 2-port with a hypothetical noiseless input termination, Z_S.

Figure 10.9 shows an equivalent circuit where the noise has been "removed" from the 2-port and assigned to the input termination by raising the temperature of the input termination to a value denoted by T_e.

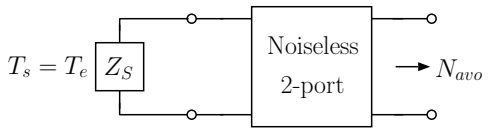


Figure 10.9: Equivalent circuit for a 2-port with a hypothetical noiseless input termination, Z_S.

T_e is the temperature which, when assigned to the input termination, produces the same available output noise power as that of the actual 2-port. In general, T_e is a function of Z_s and frequency.

T_e only characterizes the noise generated within the 2-port, not the input termination. The operating noise temperature of a system (T_{op}) is used to characterize the total amount of output noise due to the combined contributions from the 2-port and the input termination, i.e., in a practical system the source is also noisy and can be assigned a noise temperature, T_s, as in Figure 10.10. An equivalent circuit for noise analysis can be obtained by adding

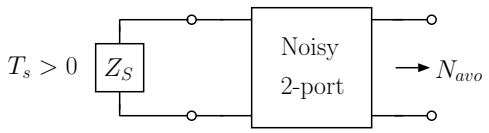


Figure 10.10: Realistic system with a noisy source.

the effective input temperature of the 2-port to the source temperature as in Figure 10.11.

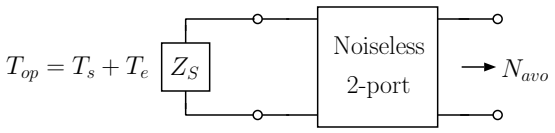


Figure 10.11: Effective input temperature of the 2-port is added to the source temperature.

T_s could represent the antenna noise or the noise from the source impedance of a signal generator. Note carefully that by adding the noise temperatures of the source and the 2-port, we have effectively added the noise powers, since noise power is proportional to temperature. The implicit assumption is that the input noise voltage due to the source and due to the 2-port are statistically independent, and hence, uncorrelated. In that case the noise powers due to the two sources add.

10.2.2 Noise Factor (F) and Noise Figure (NF)

Noise Factor (F) or *Noise Figure (NF)* are alternatives to the effective input temperature for characterizing the noise performance of a 2-port. The noise factor (F) can be defined in two equivalent ways. The first definition is based on the signal-to-noise ratio degradation between the input and output of a noisy 2-port, assuming that the noise power is calculated within the same bandwidth at both ports.

Consider Figure 10.12, which can be used as a basis for deriving an expression for the Noise Factor. The Noise Factor is defined under the assumption that the noise power per

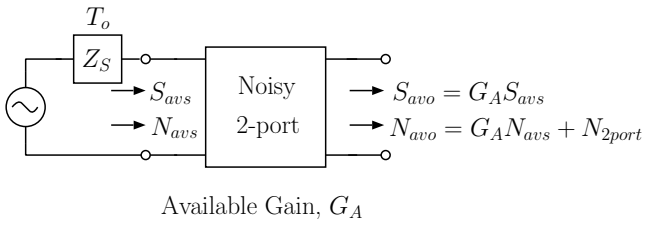


Figure 10.12: System for defining F.

unit bandwidth available from the source is constant at the value kT_o, where T_o is the standard temperature 290 K:

$$F \equiv \frac{\text{INPUT SNR}}{\text{OUTPUT SNR}}_{T_s=T_o=290K} \tag{10.22}$$

where

$$\begin{aligned} T_o &= 290\,\text{K} && \text{standard temperature} \\ S_{avs} &= && \text{signal power available from the source} \\ N_{avs} &= kT_o && \text{noise power available from the source} \\ G_A &= \tfrac{S_{avo}}{S_{avs}} && \text{available gain of the } 2-\text{port} \\ N_{2port} &= kT_eG_A && \text{noise power (1 Hz BW) available at output due to } 2-\text{port only} \end{aligned} \tag{10.23}$$

It is important to note that F does not necessarily describe the SNR degradation in a real system, because the definition is based on the assumption that the input termination has effective temperature $T_o = 290\,\text{K}$.

The second definition of F can be derived from Equation 10.22 with reference to Figure 10.12:

$$F \equiv \frac{S_{avs}/N_{avs}}{S_{avo}/N_{avo}} = \frac{S_{avs}/N_{avs}}{G_AS_{avs}/N_{avo}} = \frac{N_{avo}}{G_AN_{avs}} = \frac{G_AN_{avs} + N_{2port}}{G_AN_{avs}} \tag{10.24}$$

The final result can be thought of as an alternative definition of Noise Factor that does not make reference to any particular signal. Here F is simply the ratio of the actual noise power available at the output of the 2-port (with input termination at standard temperature) to the noise power that would be available if the 2-port was noiseless.

The noise factor can be expressed in terms of the effective input temperature of the 2-port:

$$F = 1 + \frac{N_{2port}}{G_AN_{avs}} = 1 + \frac{N_{2port}/G_A}{N_{avs}} = 1 + \frac{kT_e}{kT_o} \tag{10.25}$$

so

$$F = 1 + \frac{T_e}{T_o}, \quad T_o = 290\,\text{K}$$

Since T_o is a standard temperature (= 290 K), specification of effective input temperature (T_e) is equivalent to specifying F. For a noiseless 2-port, $T_e = 0$, which means that F = 1, i.e., there is no SNR degradation.

Noise figure (NF) is simply the noise factor expressed in dB, i.e.,

$$\begin{aligned} NF &= 10\,\log(F) \\ &= 10\,\log\left(1 + \frac{T_e}{T_o}\right) \end{aligned} \tag{10.26}$$

The relationship between NF and T_e is summarized in Table 10.1, and NF for some typical receiving systems is summarized in Table 10.2.

Noise Factor (and therefore T_e and NF) is a function of source impedance, which means that to achieve the best noise performance, the 2-port must be presented with some optimum source impedance, Z_{opt}, which may not be the impedance that gives an impedance match at the input of the 2-port — in fact, it usually is not. This means that the best NF for an amplifier doesn't always go hand-in-hand with the highest gain. See section 10.5 for more details.

NF (dB)	T_e (K)
0.25	17.2
0.50	35.4
1.0	75.1
2.0	169.6
5.0	627.1
10.0	2610

Table 10.1: Conversion between NF and effective input temperature.

System	NF (dB)
HF Comm. Receiver	6-10
VHF Comm. Receiver	1-4
Satellite Ground Station	<1
Radio Astronomy (L-Band)	<0.5

Table 10.2: Some typical Noise Figures.

10.2.3 Effective Input Temperature of a Passive Attenuator

In general, it is difficult to accurately predict the effective input temperature of a 2-port from basic physical considerations and it is necessary to measure T_e. For the special case of a passive lossy 2-port it is possible to accurately calculate the NF. (A passive lossless 2-port will have $T_e = 0$ since no noise is generated in lossless components.) An example of a passive lossy 2-port is a long piece of lossy coaxial cable.

Consider a passive 2-port network with "available loss" L ($\equiv \frac{1}{G_A} > 1$). For the moment, assume that the source termination is also passive and that the 2-port and source termination are in thermal equilibrium at *physical temperature* T_{att}. Since the network is passive and at physical temperature T_{att}, the noise power per unit bandwidth available at the output port will be $N_{avo} = kT_{att}$. This available output noise power can be thought of as the combination of the source contribution and a contribution from the 2-port. Referring the 2-port contribution to the input, representing it by an effective input temperature T_e, and equating the available output noise power to the sum of the source and 2-port contributions yields:

$$kT_{att} = kT_{att}G_A + kT_eG_A = k\frac{T_{att} + T_e}{L} \tag{10.27}$$

Solving for T_e yields:

$$T_e = T_{att}(L - 1) \tag{10.28}$$

This result is also valid when the source has an arbitrary noise temperature.

The noise factor of a passive attenuating network is:

$$F = 1 + \frac{T_e}{T_o} = 1 + \frac{T_{att}}{T_o}\,(L - 1) \tag{10.29}$$

If the physical temperature of the attenuator is T_o ($T_{att} = T_o$),

$$F = L \;\Rightarrow\; NF = 10 \log L \tag{10.30}$$

Thus, a room temperature attenuator with $10 \log L = 3\,\text{dB}$ has NF = 3 dB.

As noted in the previous section, since the available gain (and therefore, L) depends on the source impedance presented to the attenuator, T_e will depend on the source impedance.

10.2.4 Noise Temperature of Cascaded 2-ports

Consider a system consisting of a cascade of two 2-ports, as shown in Figure 10.13. Assuming

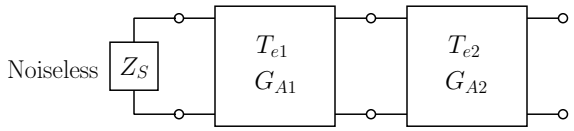


Figure 10.13: G_{A1}, G_{A2} are available power gains; T_{e1}, T_{e2} are effective input temperatures.

a noiseless source termination, the noise power (per unit bandwidth) available at the output of the first 2-port is

$$N_{avo1} = kT_{e1}G_{A1}.$$

The noise power available at the output of the second 2-port has two terms - the first is the amplified noise from the output of the first 2-port, and the second is the contribution from the second 2-port:

$$\begin{aligned} N_{avo2} &= N_{avo1}G_{A2} + kT_{e2}G_{A2} \\ &= k(T_{e1} + T_{e2}/G_{A1})G_{A1}G_{A2} \\ &= kT_eG_{A12} \end{aligned}$$

where

$$T_e = T_{e1} + \frac{T_{e2}}{G_{A1}} \tag{10.31}$$

is the effective input temperature of the cascade, and $G_{A12} = G_{A1}G_{A2}$ is the available gain of the cascade.

The Noise Factor of the cascade is:

$$F = 1 + \frac{T_e}{T_o} = 1 + \frac{T_{e1}}{T_o} + \frac{T_{e2}}{T_o}\frac{1}{G_{A1}} \tag{10.32}$$

$$F = F_1 + \frac{1}{G_{A1}}(F_2 - 1)$$

This is easily generalized to cascades of more than 2 2-ports. For a cascade of n 2-ports, the effective input temperature and Noise Factor of the cascade are given by:

$$T_e = T_{e1} + \frac{T_{e2}}{G_{A1}} + \frac{T_{e3}}{G_{A1}G_{A2}} + \ldots \frac{T_{en}}{G_{A1}G_{A2}\ldots G_{An-1}} \tag{10.33}$$

$$F = 1 + (F_1 - 1) + \frac{(F_2 - 1)}{G_{A1}} + \frac{(F_3 - 1)}{G_{A1}G_{A2}} + \ldots \frac{(F_n - 1)}{G_{A1}G_{A2}\ldots G_{n-1}} \tag{10.34}$$

Equations 10.31 through 10.34 for the cascaded noise temperature and noise factor are known as Friis' formulas.

10.2.4.1 Example - Noise temperature of cascaded amplifiers.

Find the effective input temperature, noise factor, and noise figure for the system shown in Figure 10.14.

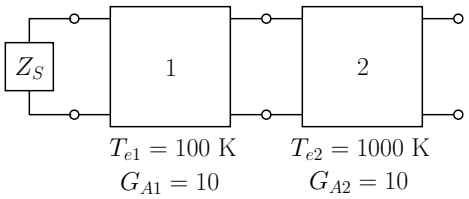


Figure 10.14: Cascaded 2-ports.

$$\begin{aligned} T_{e12} &= 100 + \frac{1000}{10} = 200\,\mathrm{K} \\ F_{12} &= 1 + \frac{200}{290} = 1.69 \\ NF_{12} &= 10 \log 1.69 = 2.3\,\mathrm{dB} \end{aligned} \tag{10.35}$$

What if the order of cascade is reversed? Assuming that the available gains, and effective input temperatures do not change, then

$$\begin{aligned} T_{e21} &= 1000 + \frac{100}{10} = 1010\,\mathrm{K} \\ F_{21} &= 4.48 \\ NF_{21} &= 6.5\,\mathrm{dB} \end{aligned} \tag{10.36}$$

This example illustrates that *the order of cascade is important* and also raises a question: for minimum cascaded noise temperature, should the 2-port with the lowest T_e (F) always come first in the cascade? The answer is no. The order of cascade that gives the best noise figure depends on the relative gains of the 2-ports, as well as their effective input temperatures. Interested readers may want to derive a formula for a figure of merit for a 2- port that depends on both the gain and effective input temperature. The figure of merit should be such that when cascading two 2-ports, the cascade has the lowest effective input temperature when the 2-port with the smaller figure of merit is used first. (The resulting figure of merit is sometimes called the *Noise Measure* of a 2-port. See homework problem 4.)

10.2.4.2 Example - Noise temperature of attenuator-amplifier cascade.

Find the effective input temperature and NF of the cascade in Figure 10.15. First, find T_e for the attenuator. A 3 dB attenuator has loss $L = 10^{3/10} \simeq 2$, thus

$$T_{e1} = T_o(L-1) = 290(2-1) = 290 \tag{10.37}$$

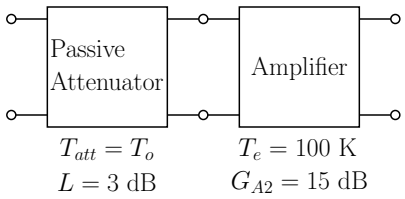


Figure 10.15: Attenuator-amplifier cascade.

The gain of the attenuator is

$$G_{A1} = \frac{1}{L} = \frac{1}{2} \tag{10.38}$$

$$\begin{aligned} \therefore T_e &= T_{e1} + \frac{T_{e2}}{G_{A1}} \\ &= 290 + \frac{100}{1/2} = 290 + 200 = \underline{\underline{490\,K}} \end{aligned}$$

$$\begin{aligned} F &= 1 + \frac{T_e}{T_o} = 1 + \frac{490}{290} = 2.69 \\ NF &= 4.3\,\text{dB} \end{aligned} \tag{10.39}$$

The NF of the amplifier was 1.3 dB, so adding the attenuator significantly increased the NF. In fact, the NF was increased by exactly 3 dB (the loss of the attenuator). The increase would not have been equal to the loss of the attenuator if the physical temperature of the attenuator was different from 290 K.

This example shows how the presence of loss in front of a low-noise amplifier can seriously degrade the overall noise temperature of the system. The example can also be worked using the noise factor. Recall that for an attenuator at standard temperature $T_o = 290\,\text{K}$

$$F_1 = L = 2 \tag{10.40}$$

and

$$F_2 = 1 + \frac{T_{e2}}{T_o} = 1 + \frac{100}{290} = 1.34 \tag{10.41}$$

$$\begin{aligned} G_{A1} &= \frac{1}{L} = \frac{1}{2} \\ F &= F_1 + \frac{(F_2 - 1)}{G_{A1}} = 2 + \frac{.34}{.5} = \underline{\underline{2.68}} \end{aligned}$$

This is the same (to within roundoff error) as the answer obtained using noise temperatures.

The preceeding example illustrates a special result that is valid for room temperature attenuation ahead of a 2-port with $NF = NF_2$, as shown in Figure 10.16.

$$NF = L(dB) + NF_2 \tag{10.42}$$

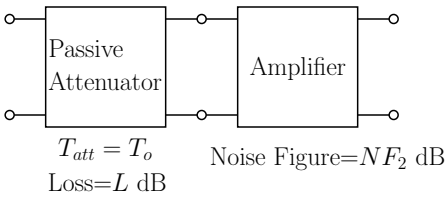


Figure 10.16: An attenuator in front of a 2-port.

i.e., the loss of the room temperature attenuator (in dB) adds directly to the NF of the second stage.

10.3 Sensitivity of a Receiving System

Sensitivity is usually specified in terms of a noise floor or, equivalently, the minimum input signal level required to give a specified SNR at the input to the demodulator. This signal level is called the minimum discernible signal, or MDS, and is specified in terms of the input power level, usually in dBm. Calculation of the SNR requires knowledge of the total noise power available from the output of the system. To facilitate calculation of the noise power, it is convenient to use the concept of equivalent noise bandwidth which is defined next.

10.3.1 Equivalent Noise Bandwidth

The available power gain of a real system may look like Figure 10.17 when plotted as a function of frequency. Usually the IF stages of a receiver will set the overall bandwidth, so the shape of this curve will reflect the shape of the IF filter's response function.

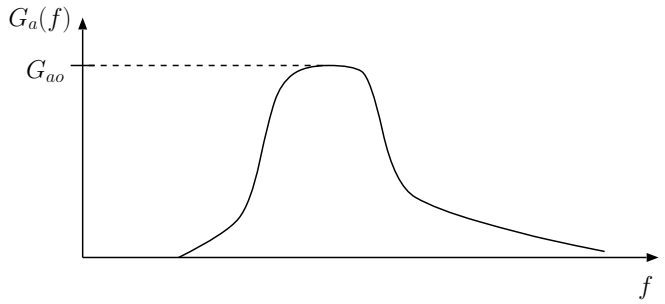


Figure 10.17: Overall available power gain of a receiver plotted as a function of frequency.

The available noise power at the output of the receiver will be

$$\begin{aligned} N_{avo} &= k\,T_{op} \int_0^\infty G_a(f)\,df \\ &= k\,T_{op}\,G_{ao} \int_0^\infty \frac{G_a(f)}{G_{ao}} df \end{aligned} \tag{10.43}$$

Equation 10.43 has the form "kTBG" if the equivalent noise bandwidth, B_n is defined as follows:

$$B_n = \int_0^\infty \frac{G_a(f)}{G_{ao}} df \tag{10.44}$$

Then Equation 10.43 becomes

$$N_{out} = k\,T_{op}\,G_{ao}\,B_n \tag{10.45}$$

In essence, by defining the equivalent noise bandwidth, the actual gain versus frequency curve is replaced with an idealized, rectangular response function as seen in Figure 10.18, where the area under the equivalent rectangular response is equal to the area under the actual response function.

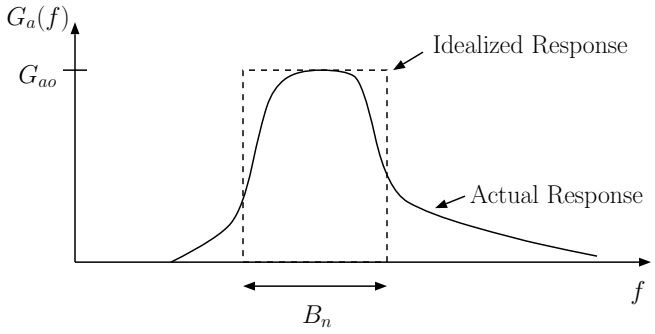


Figure 10.18: Idealized rectangular response function.

In general, the noise bandwidth of a system will be larger than the 3 dB bandwidth. For example, it can be shown that for circuits whose frequency responses are determined by simple series or parallel RLC circuits, the 3 dB bandwidth and B_n are related by:

$$B_n = \frac{\pi}{2} B_{3dB} \simeq 1.6 B_{3dB}$$

10.3.2 Noise Floor, or Minimum Discernible Signal (MDS)

As mentioned earlier, the *noise floor* or *MDS* is defined to be the input signal level required to give some specified output SNR. The following example illustrates how the noise floor is calculated.

10.3.2.1 Example - MDS for a receiving system

Find the noise floor (MDS) for the receiving system in Figure 10.19. Assume that the minimum required output signal-to-noise ratio that is required for detection of the signal is 0 dB ($SNR_{out,min} = 1$). The effective input temperature of the receiver is found as follows:

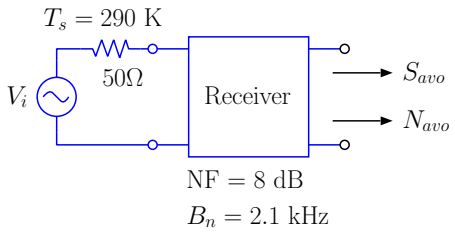


Figure 10.19: Receiving system for deriving noise floor (MDS).

$$NF = 8\,\text{dB} \tag{10.46}$$

$$F = 10^{NF/10} = 6.31 = 1 + \frac{T_e}{290}$$

$$T_e = (F-1)\,290 = 1540\,\text{K}$$

Calculate the operating noise temperature of the system:

$$T_{op} = 290K + 1540\,\text{K} = 1830\,\text{K} \tag{10.47}$$

Available output noise power:

$$N_{avo} = k\,T_{op}\,B_n\,G_a \tag{10.48}$$

Available output signal power:

$$S_{avo} = S_{avs}\,G_a \tag{10.49}$$

Output SNR:

$$SNR_{out} = \frac{S_{avo}}{N_{avo}} = \frac{S_{avs}\,G_a}{k\,T_{op}\,B_n\,G_a} \tag{10.50}$$

$$SNR_{out} = \frac{S_{avs}}{k\,T_{op}\,B_n}$$

We require $SNR_{out,min} = 1$, so the corresponding minimum available input signal power is:

$$S_{avs,min} = k\,Top B_n = (1.38)(10^{-23})(1830)(2.1)(10^3) = 5.30(10^{-17})\,\text{Watts} \tag{10.51}$$

In dBm (dB referenced to a milliwatt)

$$\begin{aligned} MDS \equiv S_{avs,min} &= 10 \log \frac{5.3\,x\,10^{-17}}{10^{-3}} \\ MDS &= -132.8\,\text{dBm}. \end{aligned} \tag{10.52}$$

Thus, the noise floor or MDS of the receiver is -132.8 dBm. Sometimes the open circuit antenna voltage is given instead of the input power. Recall that S_{avs} is defined to be the available signal power from the source. The source (antenna) impedance is 50 Ω, so

$$\begin{aligned} S_{avs,min} &= \frac{V_{s,min}^2}{8R_s} \\ &= \frac{V_{s,min}^2}{8\,(50)} \\ V_{s,min} &= 0.146\,\mu\text{V} \end{aligned} \tag{10.53}$$

This is the peak value of the open circuit antenna voltage. The rms value is

$$V_{s,min(rms)} = 0.103\,\mu\text{V} \tag{10.54}$$

Note that the output SNR used in the definition of noise floor should be interpreted as the SNR at the input to the demodulator stage. Nothing has been said about the particular type of demodulator that is used in the system. In order to specify the SNR that is required for detection of a signal, it is necessary to have some information about what SNR is required by the particular demodulator that will be used.

10.3.2.2 Example - MDS for TV receiving system

In Figure 10.20 a 300 Ω antenna is connected to a TV receiver with 300 Ω input impedance. The effective temperature of the antenna is 1000 K. The noise figure of the receiver is 4 dB. The effective noise bandwidth is 5 MHz. Find the input signal level (in dBm) required to

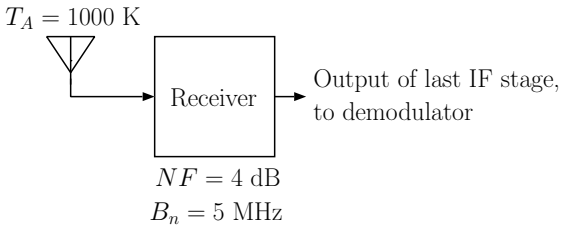


Figure 10.20: Television receiving system.

give a 15 dB signal-to-noise ratio at the output of the receiver. What is the corresponding open circuit antenna voltage?

The desired output SNR is: $SNR_{out} \Rightarrow 15dB \Rightarrow 10^{15/10} = 31.62$

$$\begin{aligned} SNR_{out} &= \frac{S_{avs}}{k\,T_{op}\,B_n} \\ S_{avs} &= k\,T_{op}\,B_n\,(SNR_{out}) \\ T_{op} &= T_s + T_e \end{aligned} \tag{10.55}$$

Find T_e from NF:

$$\begin{aligned} F &= 10^{NF/10} = 1 + \frac{T_e}{T_o} \\ F &= 2.51 \Rightarrow T_e = 438.5\,\mathrm{K} \\ T_{op} &= 1000 + 438.5 = 1438.5\,\mathrm{K} \\ S_{avs} &= 1.38 \times 10^{-23}(1438.5)(5 \times 10^6)(31.62) \\ &= 3.14 \times 10^{-12}\,\mathrm{W} = 3.14 \times 10^{-9}\,\mathrm{mW} \\ &= -85\,\mathrm{dBm}. \end{aligned} \tag{10.56}$$

Thus, -85 dBm of input signal power is required for 15 dB SNR at the output of the receiver.

The next example calculates the degradation in sensitivity resulting from addition of a lossy cable between the antenna and the receiver.

10.3.2.3 Example - TV system MDS with a lossy cable

Repeat the calculation for the case where the antenna is connected to the receiver through a long transmission line having 10 dB loss as in Figure 10.21. Calculate the effective input

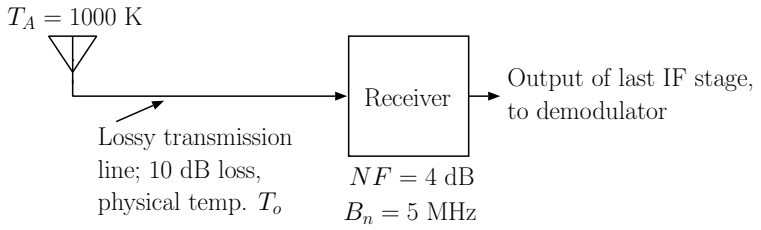


Figure 10.21: Antenna-cable-receiver.

temperature of the lossy transmission line. First convert loss from dB to power ratio: $L = 10^{10/10} = 10$. Then T_e for lossy transmission line is $T_e = T_o(L-1) = 2610$ K. Effective

input temperature of the transmission line-receiver cascade:

$$\begin{aligned} T_e &= T_{e1} + \frac{T_{e2}}{G_1} = 2610 + \frac{438.5}{1/10} \qquad (10.57) \\ &= 6995\,\mathrm{K} \quad (\mathrm{NF} = 14\mathrm{dB}) \\ T_{op} &= 1000 + 6995 = 7995\,\mathrm{K} \\ S_{avi} &= k\,T_{op}\,B_n(SNR_{out}) \\ &= 1.74 \times 10^{-11}\,\mathrm{W} = 1.74 \times 10^{-8}\,\mathrm{mW} \\ &= -77.6\,\mathrm{dBm} \end{aligned}$$

The input signal level must be 7.4 dB higher to get the same SNR as in the first example.

The next example illustrates how one can calculate the NF required of a preamp, in order to set the system MDS to some specified value.

10.3.2.4 Example - Calculating preamp NF required for a specified MDS.

In Figure 10.22 a preamplifier has been added in front of the lossy cable, in an attempt to make up for the loss in the cable. The preamp has available gain 10 dB, which counteracts the 10 dB cable loss. The question is, what NF (or T_e) must the preamp have if the overall system is to have the same MDS as the receiver alone? The NF of this system must be 4

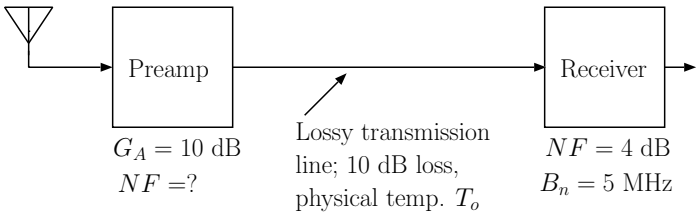


Figure 10.22: Antenna-preamp-cable-receiver.

dB if it is to be equivalent to the receiver alone.

$$\text{Preamp Gain} = 10\,\mathrm{dB} \Rightarrow G_{A1} = 10$$

$$\text{Coax Gain} = -10\,\mathrm{dB} \Rightarrow G_{A2} = \frac{1}{10}\,,\ T_{e2} = 2610\,\mathrm{K}$$

$$\text{Receiver } T_{e3} = 438.5\,\mathrm{K}$$

We want the overall $T_e = 438.5\,K$,

$$\begin{aligned} T_e &= T_{e1} + \frac{T_{e2}}{G_1} + \frac{T_{e3}}{G_1\,G_2} \qquad (10.58)\\ T_{e1} &= T_e - \frac{T_{e2}}{G_1} - \frac{T_{e3}}{G_1\,G_2} \\ &= 438.5 - 2610/10 - 438.5/(10)(1/10) \\ T_{e1} &= -261\,\mathrm{K} < 0 \end{aligned}$$

T_{e1} came out negative which means that it can't be done! Even a noiseless preamp would give an overall $T_e = 699.5\,\mathrm{K} \Rightarrow NF = 5.33\,\mathrm{dB}$! It turns out that the gain of the preamp would have to be at least 12 dB (for a noiseless preamp) and even higher for a real, noisy, preamp in order to achieve a 4dB NF for the system. This example illustrates that a preamp must more than just "make up" for the loss of an attenuator that follows it if the noise figure of the system is to be preserved.

The following section describes the principles behind practical measurements of noise temperature.

10.4 Measurement of Noise Temperature

The basic approach used for noise temperature measurements is based on a comparison of the output noise powers obtained for two different effective source temperatures. The ratio of the two resulting output noise powers can be used to determine Te and, hence, F and NF. Refer to Figure 10.23.

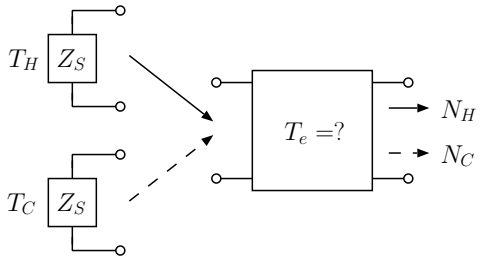


Figure 10.23: Noise temperature measurement with hot and cold loads.

$$\begin{aligned} T_H &= \text{equivalent noise temperature of the hot load} \qquad (10.59)\\ T_C &= \text{equivalent noise temperature of the cold load} \end{aligned}$$

N_H, $N_C \Rightarrow$ available output noise powers corresponding to T_H, T_C, respectively, where

$$\begin{aligned} N_H &= k(T_e + T_H)\,B_n\,G_A \qquad (10.60)\\ N_C &= k(T_e + T_C)\,B_n\,G_A \end{aligned}$$

The ratio N_H/N_C is called the "Y" factor and is the parameter that would be measured in practice:

$$Y = \frac{N_H}{N_C} = \frac{k(T_e + T_H)\, B_n\, G_A}{k(T_e + T_C)\, B_n\, G_A} \tag{10.61}$$

$$Y = \frac{T_e + T_H}{T_e + T_C}$$

Since T_H and T_C are known, it is possible to solve for T_e:

$$T_e = \frac{T_H - Y\, T_C}{Y - 1} \tag{10.62}$$

Note that measurement of the effective input temperature of a 2-port is possible without making accurate absolute power measurements. Only the ratio (relative magnitudes) of the output powers obtained with "hot" and "cold" input terminations is needed.

In commercial noise figure meters, the cold load is usually a 50 Ω resistor at room temperature. A noise diode is turned on for the hot load. The diode generates excess noise and, for the purposes of the measurement, acts like a resistor at some temperature $\gg$ 290K. The excess noise ratio is specified for noise diodes is defined by

$$ENR = 10 \log \frac{T_H - 290}{290} \tag{10.63}$$

A typical noise diode ENR is in the neighborhood of 5 dB.

When measuring small NF's the ENR must be known very accurately if precise measurements are necessary. Also, it is important that the source impedance of the noise diode doesn't change between the off (cold) and on (hot) states, because the effective input temperature and available gain of a 2-port usually depend on the source impedance.

10.4.1 Practical Considerations

In practice the NF measurement set-up looks like Figure 10.24. What is actually measured

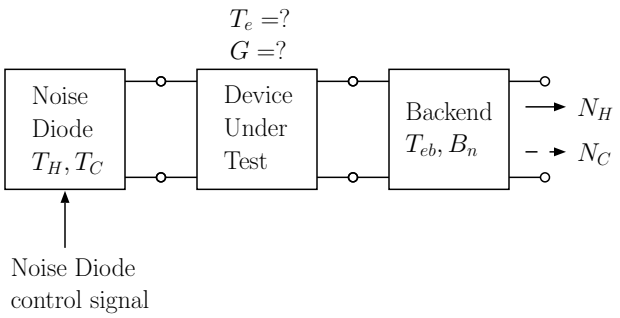


Figure 10.24: Noise Figure measurement setup.

is the effective temperature (or noise figure) of the cascade:

$$T_{ec} = T_e + \frac{T_{eb}}{G} \tag{10.64}$$

To determine T_e, it is necessary to know T_{eb} and G. These parameters can be determined if the Y-factor for the backend alone is known, as in Figure 10.25.

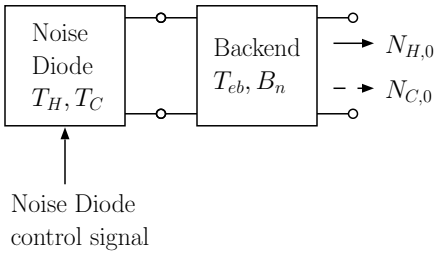


Figure 10.25: Measuring the effective input temperature of the backend.

To find the Y-factor of the backend, remove the device under test and measure the ratio of $N_{H,0}$ and $N_{C,0}$ where

$$\begin{aligned} N_{H,0} &= k(T_{eb} + T_H)\, B_n\, G_b \\ N_{C,0} &= k(T_{eb} + T_C)\, B_n\, G_b \end{aligned} \tag{10.65}$$

The Y-factor of the backend

$$Y_0 = \frac{N_{H,0}}{N_{C,0}} = \frac{T_{eb} + T_H}{T_{eb} + T_C} \tag{10.66}$$

is used to determine the effective input temperature of the backend, T_{eb}. The gain of the device under test can be determined by noticing that with the DUT in the circuit we measure

$$\begin{aligned} N_H &= k\,(T_{ec} + T_H)\, B\, G\, G_b \\ N_C &= k\,(T_{ec} + T_C)\, B\, G\, G_b \end{aligned} \tag{10.67}$$

To find G,

$$\frac{N_H}{N_{H,0}} = \frac{T_{ec} + T_H}{T_{eb} + T_H}\, G \tag{10.68}$$

Since T_{ec}, T_{eb} and T_H are known (or calculated), G can be determined. The power ratio $N_C/N_{C,0}$ could also be used to find G. In practice the measurement of the backend noise temperature (i.e., measurement of $N_{H,0}$, $N_{C,0}$) is done as part of the calibration procedure and then stored. The gain that is computed using the NF meter is averaged over the effective noise bandwidth of the backend and is therefore usually different from the gain measured at one frequency using a signal generator.

10.5 A model for the dependence of T_e on Z_S

The noise generated within a noisy 2-port can be modeled with the equivalent circuit shown in Figure 10.26 where all noise sources have been removed from the 2-port and replaced with equivalent noise voltage and noise current generators at the input of the 2-port. The voltage source accounts for all of the output noise in the case where the input of the 2-port is terminated with a short circuit and the current source accounts for all of the output noise when the input of the 2-port is left open-circuited. In general, the current and voltage

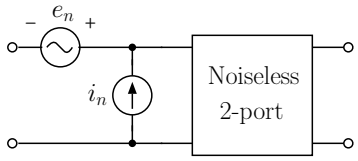


Figure 10.26: A model for noisy 2-ports.

sources must be assumed to be correlated, i.e. $< e_n i_n > \neq 0$, to account for the fact that a particular noise generation mechanism may produce output noise with both open and short circuit input terminations.

Up until now, when we have referred to noise voltages and currents, we have implicitly assumed that e_n and i_n were functions of time, i.e. that e_n and/or i_n represented the actual time-domain voltage or current waveforms that would be viewed on a instrument with some bandwidth, B. For the following discussion, it will be convenient to represent the bandlimited noise voltage and current with a complex envelope or, in other words, as *noise phasors*. For the discussion in this section, let us assume that $e_n(t)$ and $i_n(t)$ are the voltage and current waveforms in a 1 Hz bandwidth centered on some frequency, f. We can write

$$e_n(t) = \Re(E_n(t)e^{j\omega t})$$

and

$$i_n(t) = \Re(I_n(t)e^{j\omega t})$$

where $E_n(t)$ and $I_n(t)$ are complex baseband representations of the bandlimited noise signals $e_n(t)$ and $i_n(t)$. The complex signals E_n and I_n are analogous to voltage and current phasors that we are familiar with. The difference is that a purely sinusoidal signal, at some frequency, can be represented by a constant, complex, phasor whose magnitude and phase completely describe the signal. Similarly, a noise signal that has been bandlimited to a bandwidth 1 Hz, centered on some frequency, can be represented by a slowly varying[3] complex phasor. We can do circuit analysis using noise phasors in the same way that we do circuit analysis using standard signal phasors. The difference is that with noise phasors, the quantity of interest will involve the mean-square value of the noise phasor, since its mean value is always zero, and is not particularly useful. For example, given the noise phasor E_n, suppose that it is necessary to calculate the mean-square value of the corresponding time-domain noise waveform—we may do this using the following identity:

$$< e_n^2 >=< (\Re(E_n e^{j\omega t}))^2 >= \frac{< |E_n|^2 >}{2}$$

10.5.1 The relationship between T_e and input noise voltage and current.

Consider a noisy 2-port terminated with a noiseless source impedance, Z_S, as shown in Figure 10.27.

[3]The time scale for significant variation will be on the order of B^{-1}, or 1 second for $B = 1$ Hz.

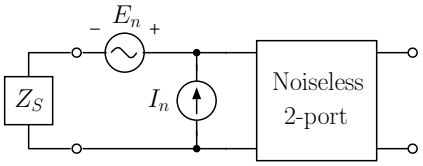


Figure 10.27: Noisy 2-port terminated with noiseless input termination.

For the purpose of deriving an expression for the effective input temperature of the 2-port, the noise generators can be associated with the source, as shown in Figure 10.28. Note that we have now switched to representing the noise voltage and current with their complex noise phasors. The effective input temperature can be found by calculating the noise power available from the source and then equating the result to kT_e.

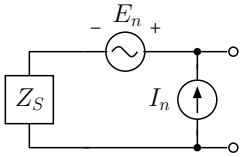


Figure 10.28: Equivalent input noise generators lumped into the source for the purpose of deriving an expression for the effective input temperature of the 2-port.

The available noise power from the source shown in Figure 10.28 is

$$kT_e = \frac{< |E_n + I_n Z_S|^2 >}{8R_S} = \frac{1}{8R_S}\{< |E_n|^2 > +2\Re\{< E_n I_n^* > Z_S\} + < |I_n|^2 > |Z_S|^2\}, \tag{10.69}$$

where the operator $\Re\{\}$ extracts the real part of its argument. Define the complex correlation coefficient, γ:

$$\gamma = \gamma_r + j\gamma_i = \frac{< E_n I_n^* >}{\sqrt{< |E_n|^2 >< |I_n|^2 >}}$$

and the noise voltage and current variances, σ_e^2, σ_i^2:

$$\sigma_e^2 =< |e_n|^2 >= \frac{< |E_n|^2 >}{2}, \quad \sigma_i^2 =< |i_n|^2 >= \frac{< |I_n|^2 >}{2}$$

Then 10.69 can be written

$$kT_e = \frac{1}{4R_S}\{\sigma_e^2 + 2\Re\{\gamma Z_S\}\sigma_e\sigma_i + \sigma_i^2|Z_S|^2\}. \tag{10.70}$$

Expand equation 10.70 using $Z_S = R_S + jX_S$ and $\gamma = \gamma_r + j\gamma_i$ and complete the squares to show that:

$$T_e = \frac{\sigma_i^2}{4kR_S}\{(R_S + \gamma_r\frac{\sigma_e}{\sigma_i})^2 + (X_S - \gamma_i\frac{\sigma_e}{\sigma_i})^2 + \frac{\sigma_e^2}{\sigma_i^2}(1 - |\gamma|^2)\}. \tag{10.71}$$

Equation 10.71 shows how the effective input temperature of a 2-port depends on the source impedance, Z_S. The particular source impedance that minimizes T_e is called the optimum source impedance for minimum noise, Z_{opt}. The corresponding minimum value of T_e is denoted by $T_{e,min}$. It is not hard to show that

$$Z_{opt} = R_{opt} + jX_{opt} = \frac{\sigma_e}{\sigma_i}\sqrt{1-\gamma_i^2} + j\gamma_i\frac{\sigma_e}{\sigma_i} \tag{10.72}$$

$$T_{e,min} = \frac{\sigma_e\sigma_i}{2k}\{\sqrt{1-\gamma_i^2} + \gamma_r\}. \tag{10.73}$$

When written in terms of Z_{opt} and $T_{e,min}$, the effective input temperature is:

$$T_e = T_{e,min} + T_o G_n \frac{|Z_S - Z_{opt}|^2}{R_S} \tag{10.74}$$

where the parameter G_n is called the noise conductance and is given by

$$G_n = \frac{\sigma_i^2}{4kT_o}. \tag{10.75}$$

For a particular 2-port, the parameters $T_{e,min}$, Z_{opt}, G_n are constants that completely characterize the noise performance of the 2-port. The set of 4 parameters $\{T_{e,min},\ Z_{opt} = R_{opt} + jX_{opt},\ G_n\}$ are called the *noise parameters* of the 2-port. Equation 10.74 shows that the effective input temperature of a 2-port increases quadratically as the source impedance moves away from the optimum value Z_{opt}. The sensitivity of T_e to changes in the source impedance is determined by the magnitude of the noise conductance parameter, G_n.

Equation 10.74 can be written in terms of reflection coefficients instead of impedances:

$$\begin{aligned} T_e &= T_{e,min} + 4G_n Z_o T_o \frac{|\Gamma_s - \Gamma_{opt}|^2}{(1-|\Gamma_s|^2)|1-\Gamma_{opt}|^2} \\ &= T_{e,min} + 4\frac{R_n}{Z_o} T_o \frac{|\Gamma_s - \Gamma_{opt}|^2}{(1-|\Gamma_s|^2)|1+\Gamma_{opt}|^2} \end{aligned} \tag{10.76}$$

where

$$\Gamma_S = \frac{Z_S - Z_o}{Z_S + Z_o} = \text{ actual source reflection coefficient}$$

$$\Gamma_{opt} = \frac{Z_{opt} - Z_o}{Z_{opt} + Z_o} = \text{ optimum source reflection coefficient}$$

$$R_n = \frac{\sigma_e^2}{4kT_o} = \text{ noise resistance}$$

In this case, the noise parameters consist of $\{R_n,\ \Gamma_{opt},\ T_{e,min}\}$ or $\{G_n,\ \Gamma_{opt},\ T_{e,min}\}$.

Contours of constant noise temperature are circles in the source reflection coefficient plane. In some cases manufacturers will provide plots of constant noise temperature (or NF) on the device data sheet. Comparison with constant available power gain curves makes it possible to select a source impedance that provides the best compromise between gain and NF.

10.5.2 Low frequency approximation and op-amp example

At lower frequencies, where op-amps are commonly employed, the noise voltage-current correlation coefficient is often ignored by assuming that it is zero, i.e. $\gamma = 0$, in which case:

$$Z_{opt} = R_{opt} = \frac{\sigma_e}{\sigma_i} \tag{10.77}$$

$$T_{e,min} = \frac{\sigma_e \sigma_i}{2k} \tag{10.78}$$

When the correlation coefficient is zero, the effective input temperature reduces to:

$$\begin{aligned} T_e &= \frac{\sigma_i^2 R_S}{4k} + \frac{\sigma_e^2}{4kR_S} \\ &= \frac{1}{2} T_{e,min} [\frac{R_S}{R_{opt}} + \frac{R_{opt}}{R_S}] \end{aligned} \tag{10.79}$$

Notice that the voltage noise dominates when R_S is small (when $R_S < R_{opt}$) and the current noise dominates when R_S is large (when $R_S > R_{opt}$). Manufacturers will provide measurements of σ_i and σ_e on the device data sheet. The rms noise voltage, σ_e, is estimated from noise measurements taken with a very small source impedance and the rms noise current is estimated from noise measurements taken with a large value of R_S. As an example, the following data was taken from the data sheet for a low-noise op-amp. At 1 kHz, the rms noise voltage and current are given as:

$$\sigma_e = 2.3\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$$

$$\sigma_i = 160\,\mathrm{fA}/\sqrt{\mathrm{Hz}}$$

The optimum source resistance for minimum noise and the associated minimum effective input temperature predicted by equations 10.77 and 10.78 are:

$$R_{opt} \simeq 14.4\,\mathrm{k\Omega}$$

$$T_{e,min} \simeq 13.3\,\mathrm{K}.$$

The minimum noise figure is therefore

$$\mathrm{NF_{min}} = 10 \log(1 + \frac{13.3}{290}) \simeq 0.2\,\mathrm{dB}.$$

It is interesting to calculate the noise temperature of this device when operated with a lower source impedance, e.g. for $R_S = 600\,\Omega$:

$$T_e = \frac{1}{2} 13.3 [\frac{0.6}{14.4} + \frac{14.4}{0.6}] \simeq 160\,\mathrm{K}.$$

In this case, operating with a source impedance of $600\,\Omega$ results in a noise temperature that is more than 10 times as large as the noise temperature with the optimum source impedance.

10.6 References

1. Krauss, H. L., C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, John Wiley & Sons, New York, 1980.

2. Smith, Jack, *Modern Communications Circuits*, McGraw Hill, 1986.

3. van der Ziel, Aldert, *Noise in Measurements*, John Wiley & Sons, New York, 1976.

4. van der Ziel, Aldert, *Noise in Solid State Devices and Circuits*, John Wiley & Sons, New York, 1986.

5. Vendelin, George D., Anthony M. Pavio, Ulrich L. Rohde, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, John Wiley & Sons, 1990.

10.7 Homework Problems

1. The input of a 2-port is terminated with a 50 Ω resistor that is at room temperature (290K), and it is found that the noise power delivered to a 50 Ω resistor at the output of the 2-port is 10^{-15} W. When the input resistor is then cooled down to 77 K (the temperature of liquid nitrogen), the output noise power drops by 2.5 dB.

   (a) Find the effective input temperature of the 2-port. You may assume that the effective input temperature is constant over the range of frequencies where the 2-port has appreciable gain.

   (b) How accurately must the change in output noise powers be measured (in dB), if the effective input temperature of the 2-port is to be measured to within plus or minus 1K?

   (c) Suppose that your measurement accuracy is plus or minus 0.1 dB. What is the uncertainty in the effective input temperature measurement?

2. A receiver has noise bandwidth $B_n = 5.0\,\text{kHz}$. The receiver has a built-in meter that gives a reading proportional to the total power at the output of the last IF stage (just before the detector). Note that the meter will indicate the total (signal + noise) power.

   Suppose that you connect a signal generator with effective source temperature $T_s = 290K$ to the input of the receiver. The meter reading is found to increase by 2 dB when the available signal power from the signal generator is raised from 0 (no output) to -130 dBm. You may assume that the noise power available from the generator is constant, and independent of the available signal power.

   What is the effective input temperature of the receiver?

3. You are given two amplifiers with the following characteristics:

$$\begin{aligned} \#1 \quad G_1 &= 4\,\text{dB} \\ T_{e1} &= 175\,\text{K} \end{aligned} \tag{10.80}$$

$$\begin{aligned} \#2 \quad G_2 &= 16\,\text{dB} \\ T_{e2} &= 200\,\text{K} \end{aligned} \tag{10.81}$$

   where G_1 and G_2 are the available gains of the amplifiers and T_{e1}, T_{e2} are the effective input temperatures. In what order should they be cascaded so that the cascade has the minimum possible noise figure? For the optimum configuration, give the effective input temperature, noise factor, and noise figure.

4. Consider a cascade of identical amplifiers with available gain G and input effective input temperature T_e.

   (a) The noise measure of an amplifier is, by definition, the effective input temperature of an infinite number of the amplifiers in cascade. Find an expression for the noise measure. Give your final result in closed form, i.e., do not leave your result in terms of an infinite series.

(b) Can the noise measure concept be applied to a 2-port with available gain that is less than 1?

(c) Suppose one is confronted with a situation like that in problem 3. That is, you are given two amplifiers with effective input temperatures T_{e1} and T_{e2} and available gains G_1 and G_2, respectively. Show that in order to achieve the lowest possible noise temperature, they should always be cascaded such that the amplifier with the lowest noise measure is in front.

5. Suppose that it is necessary to decide how to cascade N 2-ports with effective input temperatures T_i and available gains $G_i\,(>1)$, so that the cascade has the minimum possible effective input temperature. Explain how noise measure can be used to decide how to cascade the 2-ports, and why it is more efficient to use the noise measure rather than simply evaluating the cascaded noise temperature of all of the possible cascades.

6. The following situation is often encountered in practice: An antenna is located some distance from the receiver. A decision must be made whether to place the preamplifier at the antenna (before the lossy cable) or at the receiver (after the lossy cable). Suppose the distance between antenna and receiver is 100 meters. The connection will be made with a 100 meter length of coaxial cable having an attenuation of .05 dB per meter. The physical temperature of the cable will be 290K. A preamplifier is obtained that has an available gain of 16 dB and a NF of 0.50 dB.

   (a) Find the effective input temperature of the cable followed by the preamp.

   (b) Find the effective input temperature of the preamp followed by the cable.

7. The front end of a particular receiver consists of the following stages:
   Preamp:

$$\begin{aligned} G_a &= 10 \text{ dB} \\ NF &= 5 \text{ dB} \end{aligned} \tag{10.82}$$

   Mixer/LO-IF:

$$\begin{aligned} G_a &= 60 \text{ dB} \\ NF &= 11 \text{ dB} \end{aligned} \tag{10.83}$$

   (a) Suppose the preamp has an input impedance of 50 Ω and is connected to a 50 Ω antenna with effective temperature 100K. If the equivalent noise bandwidth of the system is 5 kHz, find the minimum input signal level (in dBm) required to give a 15 dB SNR at the output of the system.

   (b) Find the rms open-circuit antenna voltage (in microvolts) that will give the required SNR at the output of the system.

   (c) Now assume that the preamp noise figure is only 1 dB and repeat part 7a.

8. Consider a receiving system designed to detect signals from a deep-space probe. Suppose that the probe emits a signal with bandwidth of 100 Hz and that the receiving antenna collects enough energy to make a total signal power of -150 dBm available from the antenna. The effective antenna temperature is 30K. A preamplifier with effective input temperature of 20K and available gain of 10 dB is mounted at the antenna

terminals. The output of the preamplifier is connected to the input of the receiver through a cable with 6 dB of loss. (The physical temperature of the cable is 290K.) Assume that all ports are conjugately matched, and find the maximum effective input temperature that the receiver can have, if the signal-to-noise ratio at the detector is to be at least 3 dB. Assume that the effective noise bandwidth of the receiver is 100 Hz.

9. A spectrum analyzer is a super-heterodyne receiver with a swept LO which displays the effective input power delivered to its input terminals. The analyzer computes the effective input power by measuring the total (signal+noise) power delivered by the last IF stage and dividing this power by the total power gain of all the preceding stages.

   Suppose that you want to measure the effective input temperature of a 2-port using a spectrum analyzer and you know that the available gain of the 2-port is 25 dB. You may assume that the input and output of the 2-port are conjugately matched when terminated in 50 Ω, and that the input impedance of the spectrum analyzer is 50 Ω.

   With a room temperature (290K) 50 Ω load connected directly to the input of the spectrum analyzer, the analyzer indicates that the effective input noise power is -168 dBm in a bandwidth of 1 Hz centered on the frequency of interest. You now connect the 50 Ω load to the input of the 2-port and connect the output of the 2-port to the spectrum analyzer. With this configuration you find that the equivalent input noise power in a 1 Hz bandwidth measured by the spectrum analyzer is -140 dBm. Remember that the number displayed by the spectrum analyzer is the equivalent input power at the analyzer's input port and not at the input of the 2-port under test.

   (a) What is the effective input temperature of the spectrum analyzer?

   (b) Find the effective input temperature of the 2-port.

10. The signal distribution network for cable television systems consists of cascaded sections of lossy transmission line and amplifiers. Suppose that the lossy cable and amplifiers are both matched to 75Ω. Each lossy transmission line has loss 0.05 dB/meter, length 500 meters and **physical temperature** 290K. Each amplifier has available gain 25dB and effective input temperature $T_e = 100$K. Note that the gain of the amplifier is just enough to make up for the loss in the cable, i.e., the cascade of a lossy t-line and an amplifier results in a new 2-port with 0dB gain.

    Now suppose that it is necesssary to distribute a television signal from a main distribution station to a remote site. The site is 1.5km from the main station, so it will be necessary to use 3 sections of t-line and 3 amplifiers.

    (a) Suppose that the t-lines and amplifiers are cascaded in the order:

    $$\mathrm{t-line} \Rightarrow \mathrm{amplifier} \Rightarrow \mathrm{t-line} \Rightarrow \mathrm{amplifier} \Rightarrow \mathrm{t-line} \Rightarrow \mathrm{amplifier}$$

    Find the effective input temperature of the cascade of all of the amplifiers and cables.

    (b) Suppose that the t-lines and amplifiers are cascaded in the order:

    $$\mathrm{amplifier} \Rightarrow \mathrm{t-line} \Rightarrow \mathrm{amplifier} \Rightarrow \mathrm{t-line} \Rightarrow \mathrm{amplifier} \Rightarrow \mathrm{t-line}$$

    Find the effective input temperature of the cascade of all of the amplifiers and cables.

(c) Suppose that the source at the main station has equivalent source temperature $T_S = 290$K. Assuming that the available signal power and system bandwidth is the same in both cases, specify the improvement (in dB) that would result in the output SNR if cascade (10a) were changed to cascade (10b).

11. Find the equivalent noise bandwidth, B_n, in terms of the -3 dB bandwidth, B_{3dB}, for a system with available gain function equivalent to a lowpass Butterworth response with order n. In other words, assume that the available gain function has the following form:

$$G_A(f) = \frac{1}{1 + (\frac{f}{B_{3dB}})^{2n}}$$

Give results for $n = 1, 2, 3, 4$.

12. Derive equations 10.76 starting with equation 10.74.

Chapter 11

Mixers

Mixers may be classified according to whether they are based on active or passive devices. Another distinction, that can apply to either passive or active mixers, is whether mixing occurs as a result of a *soft* nonlinearity such as the current-voltage relationship in a diode or transistor, or whether mixing results from a *hard* nonlinearity such as from a switch. Most mixers in use today are of the *switching type*, whereby diodes or transistors are used to switch the connection between the RF input and the IF output at a rate that is controlled by the local oscillator. We will give a brief overview of several representative mixer circuits in this chapter.

11.1 Mixers Based on Gradual Nonlinearities

11.1.1 Single-ended BJT Mixer

The single-ended BJT mixer makes use of the nonlinear relationship between the base-emitter voltage and the collector current. If the RF and LO voltages are applied to the base of the BJT so that the base-emitter voltage $v_i(t)$ has both RF and LO signal components then the collector current will contain terms that are proportional to all powers of the total input voltage:

$$i_C = i_{DC} + k_1\, v_i(t) + k_2\, v_i^2(t) + k_3\, v_i^3(t) + \ldots \tag{11.1}$$

In general, this produces all possible mixing products, including:

$$f_{RF},\ f_{LO},\ |f_{RF} \pm f_{LO}|,\ |2f_{RF} \pm f_{LO}|,\ |2f_{LO} \pm f_{RF}|, \ldots \tag{11.2}$$

The output voltage can be developed by loading the collector with a parallel resonant circuit so that a significant output voltage is developed by only one of the frequency components present in the collector current waveform. The collector circuit is usually tuned to resonate at either $f_{RF} + f_{LO}$ or $|f_{RF} - f_{LO}|$. Figure 11.1 is an example of a single-ended BJT mixer.

In a practical implementation of the mixer shown in Figure 11.1, a matching network would be used between the RF source and the RF input in order to couple the maximum amount of RF signal to the base of the transistor. The matching network would be designed to present a high impedance to the LO signal. Similarly, a matching network would be used between the LO source and the LO input, and this network would be designed to present a high impedance to the RF source. In this way, the LO source is prevented from loading the

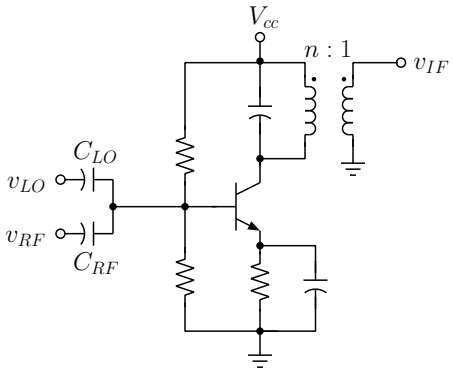


Figure 11.1: BJT mixer

RF source, and vice versa. A simpler compromise is to supply the LO voltage through a small-value coupling capacitor (C_{LO}) such that the LO source does not significantly disturb the matching for the RF signal. For this circuit the local oscillator drive level required is on the order of 100mV. A BJT mixer of this type has conversion gain $\simeq g_m R_L$ where R_L is the load resistance seen by the collector at f_{IF}. This circuit is attractive because of its simplicity, and it has been used in inexpensive mass-produced receivers for consumer applications. It is not a high performance mixer, however. In general, a single-ended BJT mixer will have a higher noise figure than a mixer that employs an FET, or a properly designed passive mixer based on a diode bridge. In addition, BJT mixers are subject to severe intermodulation distortion. To avoid this problem, v_{RF} must be kept smaller than ~ 10 mV. In general, BJT mixers of this type have poor large signal characteristics.

11.1.2 Single-ended FET Mixers

If a FET is operated in its constant current region, then

$$i_D = I_{DSS}(1 - \frac{v_{gs}}{V_p})^2 \tag{11.3}$$

The circuit is similar to the one shown in Figure 11.1 for the BJT mixer, but the FET has two advantages over the BJT. These are:

- a much lower third-order IMD, since there is no cubic term in the i_D versus v_{gs} relationship
- Much higher RF input voltages are usable, i.e., up to at least 100 mV.

A disadvantage of the FET mixer is a somewhat smaller conversion gain than the BJT circuit.

A major drawback of the single-ended BJT and FET mixers is the presence of a local oscillator component as well as an RF component at the output of the mixer. Although the LO and RF signals will be filtered by the tuned output circuit, some of the relatively large LO signal will inevitably leak through the mixer to the IF stages. A large LO component

at the input to the IF stages is obviously undesirable, since it can cause overload at the IF stages. It can also create undesirable effects such as gain compression and possible generation of intermodulation products in the IF stages.

11.1.3 Balanced Mixers

A balanced configuration can be used to effectively remove the LO and/or RF signals from the output of the mixer. One type of balanced mixer often used in integrated circuits is the differential-pair multiplier as shown in Figure 11.2.

The multiplier is based on a differential amplifier. For a single transistor biased in the

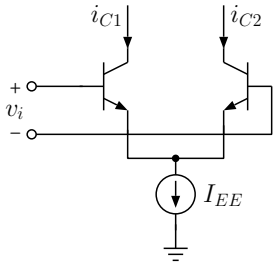


Figure 11.2: A differential-pair amplifier. The transistors are assumed to be biased in the active region. The bias network is not shown.

active mode

$$i_C = I_S e^{v_{BE}/V_T}. \tag{11.4}$$

For the emitter-coupled pair with identical transistors ($I_{S1} = I_{S2}$, $V_{T1} = V_{T2}$)

$$\frac{i_{C1}}{i_{C2}} = e^{v_i/V_T} \tag{11.5}$$

where

$$v_i = v_{BE1} - v_{BE2}$$

Since the emitters are connected to a constant current source,

$$i_{E1} + i_{E2} = I_{EE} \tag{11.6}$$

or

$$\frac{1}{\alpha}(i_{C1} + i_{C2}) = I_{EE} \tag{11.7}$$

From Equation 11.5

$$i_{C1} = i_{C2} e^{v_i/V_T} \tag{11.8}$$

so using Equation 11.7

$$i_{C1} = \frac{\alpha I_{EE}}{1 + e^{-v_i/V_T}} \tag{11.9}$$

$$i_{C2} = \frac{\alpha I_{EE}}{1 + e^{v_i/V_T}} \tag{11.10}$$

If the output is taken between the collectors of the two transistors as in Figure 11.4, then the output voltage will be proportional to the difference of the emitter currents:

$$\begin{aligned} i_{C1} - i_{C2} &= \alpha I_{EE} \frac{e^{v_i/2V_T} - e^{-v_i/2V_T}}{e^{v_i/2V_T} + e^{-v_i/2V_T}} \\ \Delta i_C &= \alpha I_{EE} \tanh \frac{v_i}{2V_T} \end{aligned} \tag{11.11}$$

A plot of the current difference, $\Delta i_C = i_{C1} - i_{C2}$, is shown in Figure 11.3. If the input

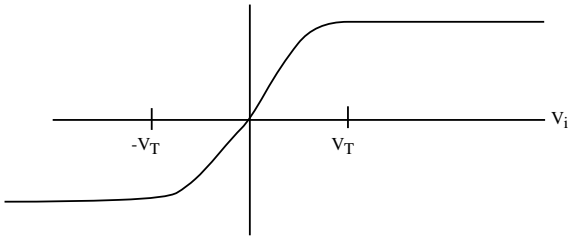


Figure 11.3: Current difference

signal levels are small, i.e., if $\frac{v_i}{2V_T} \ll 1$, then the hyperbolic tangent can be approximated as follows:

$$\tanh\left(\frac{v_i}{2V_T}\right) \simeq \frac{v_i}{2V_T} \tag{11.12}$$

Thus

$$\Delta i_C \simeq \alpha I_{EE} \frac{v_i}{2V_T} \tag{11.13}$$

Note that I_{EE} multiplies the input signal in equation 11.13. If a second signal, i_{i2}, is added to I_{EE} then the output will contain a term that is proportional to $v_i i_{i2}$, i.e. let $I_{EE} \rightarrow I_{EE} + i_{i2}$, the the output is

$$\Delta i_C \simeq \frac{\alpha}{2V_T}(I_{EE} + i_{i2})v_i \tag{11.14}$$

Figure 11.4 shows a balanced mixer based on the differential amplifier. The output voltage is $v_o = \Delta i_c R$. The tail current for the differential pair is $I_{EE} + i_{i2}$, where i_{i2} represents the time-varying current that results from the input voltage signal v_{i2}.

According to equation 11.14, the multiplier based on the differential pair yields an output that is proportional to $(K + v_i)v_{i2}$. Hence the output contains the desired product term plus the input signal v_{i2}. Suppose that a balanced multiplier of the type shown in Figure 11.4 is used with the RF signal driving the differential-amplifier input ($v_i \rightarrow v_{RF}$) and the LO signal driving the tail-current source ($v_{i2} \rightarrow v_{LO}$). In that case the output will be proportional to $v_{RF}(K + v_{LO}) = Kv_{RF} + v_{RF}v_{LO}$. Notice that the RF signal is present in the output, along with the desired product signal. On the other hand, if $v_i \rightarrow v_{LO}$ and $v_{i2} \rightarrow v_{RF}$, then the output is proportional to $v_{LO}(K + v_{LO}) = Kv_{LO} + v_{RF}$, and the LO signal accompanies the desired product signal.

The unwanted term can be eliminated as follows—suppose that we construct two identical multipliers, and drive the differential inputs of each amplifier with v_i. Drive the tail-current input of amplifier 1 with v_{i2} and the tail-current amplifier of multiplier 2 with $-v_{i2}$.

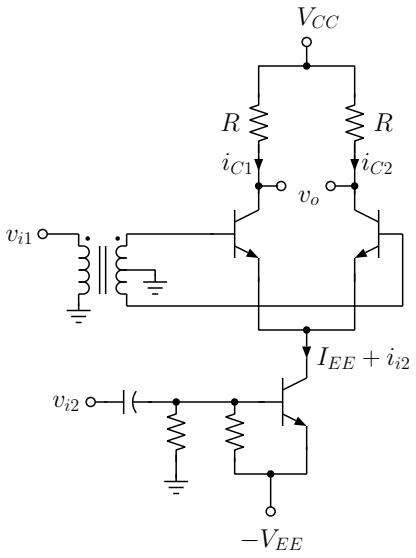


Figure 11.4: Differential-pair multiplier circuit.

The output of multiplier 1 is then proportional to $Kv_{i2} + v_i v_{i2}$, whereas the output of multiplier 2 is proportional to $Kv_{i2} - v_i v_{i2}$. The difference between the two outputs will contain only a term proportional to $v_i v_{i2}$ because the unwanted term will be canceled out. Multipliers based on this principle are called *Gilbert cell* multipliers, and are named after Barrie Gilbert who described a practical, and precise, multiplier circuit based on the technique.[1] Notice that in order to obtain an output that contained only the desired product term, and no RF or LO term, it is necessary to use a double-balanced configuration, where both the LO and RF inputs are applied differentially.

11.2 Mixers Based on Switches

Most mixers used today are based on switches. The basic idea can be illustrated using a very simple circuit as in Figure 11.5. Suppose that the switch is closed when $v_{LO} > 0$ and open when $v_{LO} < 0$. In addition, let $v_{RF}(t) = V_{RF} \cos \omega_{RF} t$ and $v_{LO}(t) = V_{LO} \cos \omega_{LO} t$, then

$$v_o(t) = V_{RF} \cos \omega_{RF} t \; p(t) \tag{11.15}$$

where p(t) is a switching function. The switching function for this circuit is shown in Figure 11.6.

The switching function can be expanded in a Fourier series:

$$p(t) = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{\sin \, n\pi/2}{n\pi/2} \cos \, n\, \omega_{LO} t \tag{11.16}$$

[1]"A Precise Four-Quadrant Multiplier with Subnanosecond Response," by Barrie Gilbert, *IEEE Journal of Solid-State Circuits*, Vol. SC-3, No. 4, December 1968, p 265.

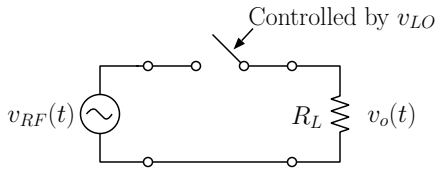


Figure 11.5: Switching principle

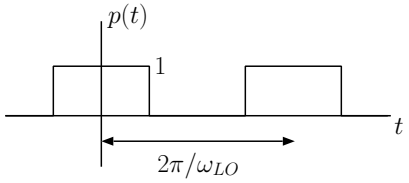


Figure 11.6: Switching function.

Using the Fourier series respresentation for $p(t)$ the output voltage can be written as

$$v_o(t) = V_{RF} \cos \omega_{RF} t \left\{ \frac{1}{2} + \sum_{n=1}^{\infty} \frac{\sin n\pi/2}{n\pi/2} \cos n\,\omega_{LO} t \right\} \tag{11.17}$$

Notice that the coefficient $\frac{\sin n\pi/2}{n\pi/2}$ is equal to 0 when n is an even number. Therefore, the output signal has components at the following frequencies:

$$f_{RF},\ |f_{RF} \pm n\ f_{LO}| \quad n = 1, 3, 5, \ldots \tag{11.18}$$

A filter can be used to select the desired component which is usually $|f_{RF} \pm f_{LO}|$. It is also possible to select $|f_{RF} \pm n\,3f_{LO}|$ or higher order terms - this is called harmonic mixing.

There are many ways to implement a switch that is controlled by the local oscillator signal. Diodes are often used as switches. If a relatively large LO voltage is impressed on a diode, the diode will be forward biased on positive excursions of the LO signal, and will be in a low-impedance state for superposed, small RF signals. On negative excursions of the LO voltage the diode is driven into a high impedance state and behaves like an open switch. Transistors can also be used as switches, with the LO controlling the bias current and determining whether the device is biased into a low-impedance state, or is cutoff.

The differential multiplier circuit described in the previous section is often used to implement a switching mixer. Refer back to Figure 11.4 and suppose that $v_i \to v_{LO}$. Allow the amplitude of v_{LO} to be *large* compared to V_T; then the $\tanh \frac{v_{LO}}{2V_T}$ term will be driven to $+1$ on positive excursions of v_{LO} and to -1 on the negative excursions. In this mode of operation, the transistors are driven as switches, as shown in Figure 11.7. The output voltage for the switch configuration shown is $v_o = -(I_{DC} + i_{RF})R$. When v_{LO} changes sign, the states of the switches reverse, and the output voltage is $v_o = +(I_{DC} + i_{RF})R$. The output can be written as

$$v_o(t) = -(I_{DC} + i_{RF})\ R\ p(t)$$

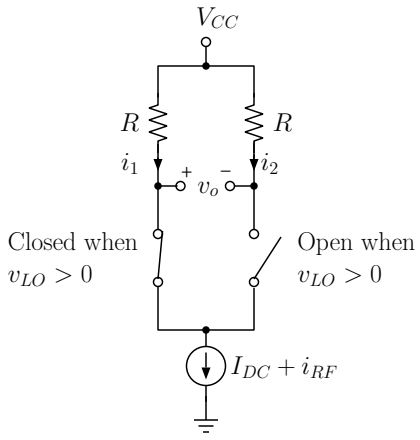


Figure 11.7: Equivalent circuit for differential pair when a large LO signal is applied to the differential amplifier.

where $p(t)$ is the symmetrical switching function shown in Figure 11.8.

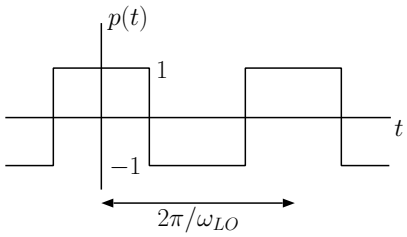


Figure 11.8: Symmetrical switching function

This switching function has the following Fourier series:

$$p(t) = 2\sum_{n=1}^{\infty} \frac{\sin n\pi/2}{n\pi/2} \cos n\omega_{LO} t. \tag{11.19}$$

Writing $i_{RF}(t) = I_{RF} \cos \omega_{RF} t$:

$$v_o(t) = -(I_{DC} + I_{RF} \cos \omega_{RF} t) \left\{ 2 \sum_{n=1}^{\infty} \frac{\sin n\pi/2}{n\pi/2} \cos n\omega_{LO} t \right\} \tag{11.20}$$

The frequency components at the output will be

$$nf_{LO}, \quad |f_{RF} \pm nf_{LO}| \quad n = 1, 3, 5, \ldots \tag{11.21}$$

Two differential switching pairs can be combined in a double-balanced *Gilbert cell* arrangement to cancel the unwanted LO component. This is illustrated in Figure 11.9. Sup-

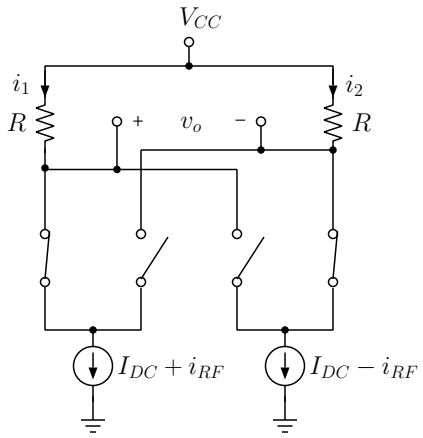


Figure 11.9: Two differential switching pairs combined in a double-balanced configuration.

pose that the switches are in the state shown when $v_{LO} > 0$ and that all switches change state when $v_{LO} < 0$. Then, the currents are $i_1 = I_{DC} + p(t)i_{RF}$ and $i_2 = I_{DC} - p(t)i_{RF}$. The output voltage is $v_o = R(i_2 - i_1) = -2Ri_{RF}(t)p(t)$, hence the output contains only components at the frequencies $|f_{RF} \pm nf_{LO}|$. Mixers of this type are often implemented in integrated circuits.

The diode-ring double-balanced mixer shown in Figure 11.10a, b is another circuit that produces an output signal proportional to the product of the RF input signal and a switching function, i.e. $v_{RF}(t)p(t)$. This mixer is noted for its ability to tolerate large RF signals with relatively small nonlinear distortion, and for its relatively low noise figure. Its opera-

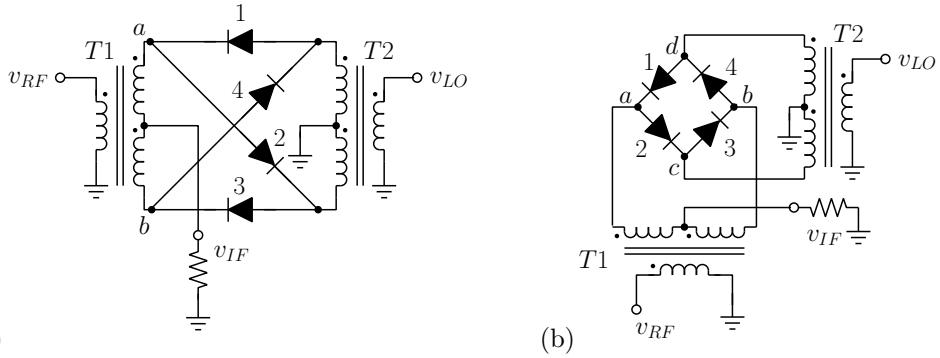


Figure 11.10: Four-diode double-balanced mixer. Figures (a) and (b) are the same circuit. Figure (b) emphasizes the fact that the mixer is based on a diode "ring" topology.

tion is easily understood once the operation of the 3-winding transformers (voltage baluns) is understood. For simplicity, let's assume that the transformers are *ideal* 3-winding trans-

formers. If the voltage across the windings is denoted by V_A, V_B, and V_C, with polarity indicated by the dots, and the current in each winding is denoted by i_A, i_B, and i_C, with positive current defined to flow into the dot, then an *ideal* 3-winding transformer satisfies

$$V_A = V_B = V_B, \qquad i_A + i_B + i_C = 0.$$

Physically, these relationships arise because when three windings with the same number of turns are tightly coupled, e.g. by winding all three on a high-permability ferrite toroid, then the magnetic flux through all 3 windings will be the same and the emf induced in each winding will be the same. Since the ideal transformer is a lossless device, power conservation requires that the sum of the current flowing out of (or into) the dots of all windings must be zero. In the context of the circuit described here, if current i flows into the dot of the primary (driven) winding, the sum of the currents flowing out of the dots of the secondary windings must be i.

We can now describe the operation of the 4-diode ring mixer. The three-winding transformers are assumed to be ideal, hence the voltage impressed across the primary will be reflected across both secondary windings, with polarity indicated by the dots. Refering to Figure 11.10b, when v_{LO} is positive, the voltage at the top of $T2$'s secondary is positive and the voltage at the bottom of $T2$'s secondary is negative. Therefore diodes 1 and 2 are "on" and in a low-impedance state as far as the small RF signal is concerned. Diodes 3 and 4 will be "off", or in a high-impedance state, so the right-hand side winding of transformer $T1$'s secondary is effectively disconnected from the system. Refer to Figure 11.11a where the forward biased diodes have been replaced with an incremental resistance, r_d, and the reverse biased diodes have been removed, as they are assumed to be effectively "open circuits". The symmetry of the circuit ensures that the voltage v_{RF} across the left-hand winding of $T1$'s secondary will drive equal currents through diodes 1 and 2. These currents flow to ground through the windings of $T2's$ secondary. Half of the RF current flows into the dot on the top of $T2$'s secondary and the other half flows into the bottom of $T2$'s secondary. These currents produce opposing magnetic fluxes and no voltage is developed across the windings of $T2$ secondary. Therefore, the RF voltage at nodes d and c is zero, and nodes c and d are *virtual grounds* for the RF signal. The RF voltage applied by the left-hand winding of $T1$'s secondary appears across the series combination of the load resistance R_L and the parallel combination of diodes 1 and 2 ($r_d/2$). The voltage across the load is $v_{IF} = -v_{RF}\frac{R_L}{R_L+r_d/2}$. When the polarity of v_{LO} reverses so that the voltage at the dots of $T2$'s secondary is negative, the situation reverses, i.e. diodes 1 and 2 are "off" and diodes 3 and 4 are "on". Nodes c and d are still virtual grounds for the RF signal, so the voltage across the load resistor is $v_{IF} = v_{RF}\frac{R_L}{R_L+r_d/2}$. If the forward resistance of the diodes is small compared to the load resistor ($r_d \ll R_L$), then the output from the double-balanced 4-diode ring mixer can be written as

$$v_{IF}(t) = -v_{RF}(t)p(t)$$

where $p(t)$ is the symmetrical switching function shown in Figure 11.8. Since the switching function has no DC component, there is no direct leakage of the RF signal into the IF output. Also, by virtue of the balance in the circuit, the voltage at nodes a and b due to the LO signal is always zero so there is no leakage of the LO signal back into $T1$, i.e. LO-RF isolation is high.

For diode-based switching-type mixers to work as described above, it is important that the local oscillator signal (v_{LO}) be large, so that the diodes will be biased to a low impedance state, and so that it can be assumed that v_{LO} controls the state of the diodes at all times.

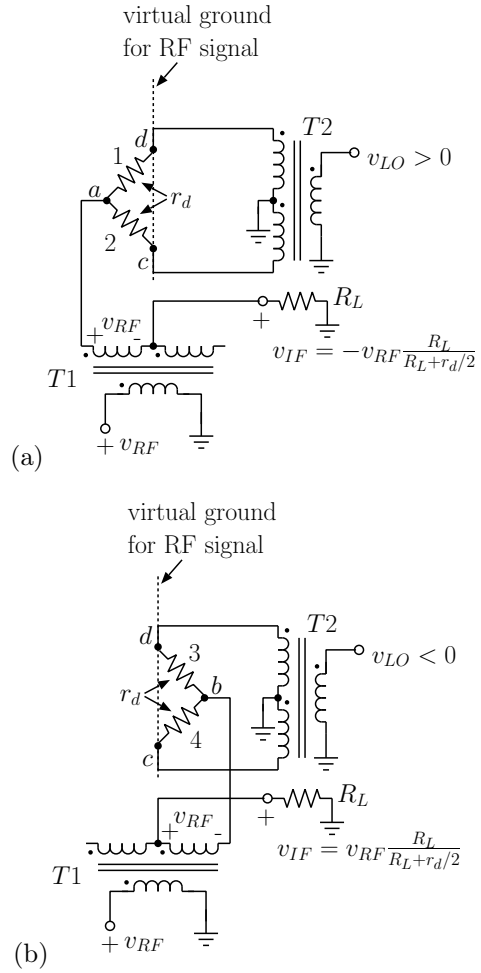


Figure 11.11: 4-diode ring double-balanced mixer equivalent circuit when (a) $v_{LO} > 0$ and (b) $v_{LO} < 0$.

Thus passive mixers require relatively high LO drive levels. Typically, the peak value of the LO signal must be at least 0.7V in order to turn on the diodes completely. Assuming that the input impedance seen by the LO source is 50 Ω, the required input level is approximately 7 dBm. The harmonic mixing components, $|mf_{RF} \pm nf_{LO}|$, can cause spurious responses in a receiver in addition to the usual image response. These responses must be carefully considered when designing a receiver. This will be addressed in section 11.4.

11.3 Conversion Loss in Mixers

Conversion loss is defined to be

$$L_C = 10 \log_{10} \frac{P_{in}}{P_{out}}$$

where P_{in} is the power delivered in to the mixer at the RF frequency, and P_{out} is IF output power delived to the load at the desired output frequency (usually $|f_{RF} \pm f_{LO}|$). Let's take the double-balanced diode-ring mixer as an example. Assuming that the diode forward resistance is small compared to the load resistance, the output of the mixer is

$$v_{IF}(t) = -V_{RF} \cos \omega_{RF} t \left\{ 2 \sum_{n=1}^{\infty} \frac{\sin \frac{n\pi}{2}}{\frac{n\pi}{2}} \cos n\omega_{LO} t \right\}.$$

If a filter selects just one output component, then the peak amplitude of the output voltage will be

$$V_{IF} = V_{RF} \left| \frac{\sin \frac{n\pi}{2}}{\frac{n\pi}{2}} \right|.$$

In an ideal 4-diode mixer, the RF input is always "looking" through the transformer primary to one of the secondary windings and through the switches to the load. If the switches are lossless, then, the RF input impedance will be equal to the load impedance. Hence, the RF input voltage and the IF output voltage are developed across the same impedance, so the ratio of P_{in} and P_{out} is the same as the ratio $(V_{RF}/V_{IF})^2$. The conversion loss, for any n, is therefore

$$L_C = 20 \log_{10} \left| \frac{n\pi}{2 \sin \frac{n\pi}{2}} \right|.$$

The conversion loss is tabulated in Table 11.1. When the desired term is $|f_{RF} \pm f_{LO}|$, the conversion loss for the ideal switching mixer is 3.9 dB. In practice, losses in the transformer and diodes usually raise this value to 5-6 dB.

n	Ouput Frequency	Loss (dB)
1	$\|f_{RF} \pm f_{LO}\|$	3.9
3	$\|f_{RF} \pm 3f_{LO}\|$	13.5
5	$\|f_{RF} \pm 5f_{LO}\|$	17.9

Table 11.1: Conversion loss for ideal double-balanced diode-ring mixer.

11.4 Spurious Responses in Receivers - Spur Charts

The idealized models for double-balanced switching mixers predict that the output will contain components at frequencies given by $|f_{RF} \pm nf_{LO}|$, with n an odd integer. When more realistic models for the switches are considered, it is found that harmonics of the RF signal will be generated within the mixer, and that even-order harmonics of the LO will be present. A fairly general model for the nonlinear properties of a mixer consists of a nonlinear 2-port with non-zero coefficients k_1, k_2, k_3, ... in front of an ideal multiplier which is driven by a local oscillator signal containing the fundamental frequency component ω_{LO} in addition to harmonics of the local oscillator signal, $2\omega_{LO}$, $3\omega_{LO}$, ... etc.

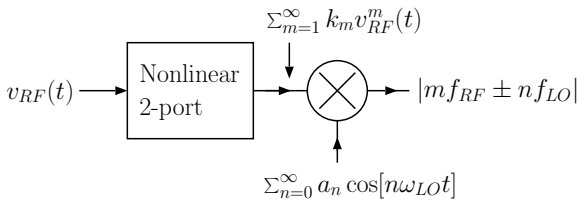


Figure 11.12: A model for nonlinearities in a mixer. The multiplier shown in the figure is an ideal multiplier. The output of the mixer will include terms resulting from mixing between all harmonics of the RF and LO signals. Note that the $n = 0$ (DC) component is included in the output of the LO in order to model the direct leakage of harmonics of V_{RF} through the mixer to the IF port.

When evaluating a particular receiver design, it is highly desirable to consider the possible spurious responses which will occur because of the non-ideal characteristics of the mixer. The goal of this section is to determine the frequencies of all possible input signals which could be mixed to the IF.

The receiver will be designed such that the desired signal frequency, f_D, is related to the IF, f_{IF}, by one of the following relationships:

$$f_{LO} = f_D + f_{IF} \tag{11.22}$$

or

$$f_{LO} = f_{IF} - f_D \tag{11.23}$$

or

$$f_{LO} = f_D - f_{IF} \tag{11.24}$$

Equation 11.22 corresponds to the choice of "high LO" for either the up- or down-conversion receivers. Equation 11.23 corresponds to "low LO" for the up-conversion receiver, and equation 11.24 corresponds to "low LO" for the down-conversion receiver.

A spurious signal that mixes to f_{IF} satisfies the following equation:

$$|mf_S \pm nf_{LO}| = f_{IF} \tag{11.25}$$

where we assume that m, n are integers; $m > 0$, $n \geq 0$. Notice that by allowing $n = 0$ we allow for possible IF responses due to direct leakage of the RF signal (or its harmonics) into

the IF passband. Taking account of the absolute value sign, equation 11.25 can be expanded into 3 equations:

$$\begin{aligned} mf_S + nf_{LO} &= f_{IF} \\ mf_S - nf_{LO} &= f_{IF} \\ mf_S - nf_{LO} &= -f_{IF} \end{aligned} \tag{11.26}$$

Now, depending on the receiver configuration of interest, one of equations 11.22, 11.23, or 11.24 may be used in equations 11.26 to eliminate f_{LO} and to solve for the frequency of possible spurious signals as a function of the receiver tuning, represented by the desired signal frequency f_D. Notice that in this context, f_D may be interpreted as the receiver's "dial setting" since it is the number that the receiver display will indicate. In some literature f_D is referred to as the "tuned" frequency, since it is the intended frequency that the receiver is tuned to receive.

Suppose that the case of interest is "high LO". Then, using equation 11.22 in 11.26 we find:

$$\begin{aligned} mf_S &= (1-n)f_{IF} - nf_D \\ mf_S &= (1+n)f_{IF} + nf_D \\ mf_S &= (n-1)f_{IF} + nf_D \end{aligned} \tag{11.27}$$

Since the right-hand side of the first and third equations in 11.27 are the same except for a factor of -1, and we are interested in positive spurious frequencies, the first equation is not necessary and the relevant equations are:

$$\begin{aligned} mf_S &= (n-1)f_{IF} + nf_D \\ mf_S &= (n+1)f_{IF} + nf_D \end{aligned} \tag{11.28}$$

It is convenient to normalize all frequencies by the IF, i.e. define:

$$S \equiv \frac{f_S}{f_{IF}}, \text{ and } D \equiv \frac{f_D}{f_{IF}}$$

then we obtain the equations used to generate a so-called "universal spur chart" for any receiver employing "high LO":

$$\begin{aligned} S &= \tfrac{n-1}{m} + \tfrac{n}{m}D \\ S &= \tfrac{n+1}{m} + \tfrac{n}{m}D \end{aligned} \tag{11.29}$$

For each value of m and n these equations define two lines, each of which determines the normalized spur frequency, S, for a particular normalized tuning frequency D. The case $m = 1, n = 1$ gives the lines $S = D$ and $S = D + 2$ which correspond to the desired frequency and the regular image, respectively. These responses were discussed already in Chapter 3. The other lines correspond to spurious responses that result from the non-ideal nature of the mixer.

The normalized spur frequency is plotted in Figure 11.13 for all possibilities up through "fifth order" where the "order" of a spur is defined to be the value of $n+m$. The m, n values for each curve are plotted on each line. The part of the plot where $0 < D < 1$ corresponds

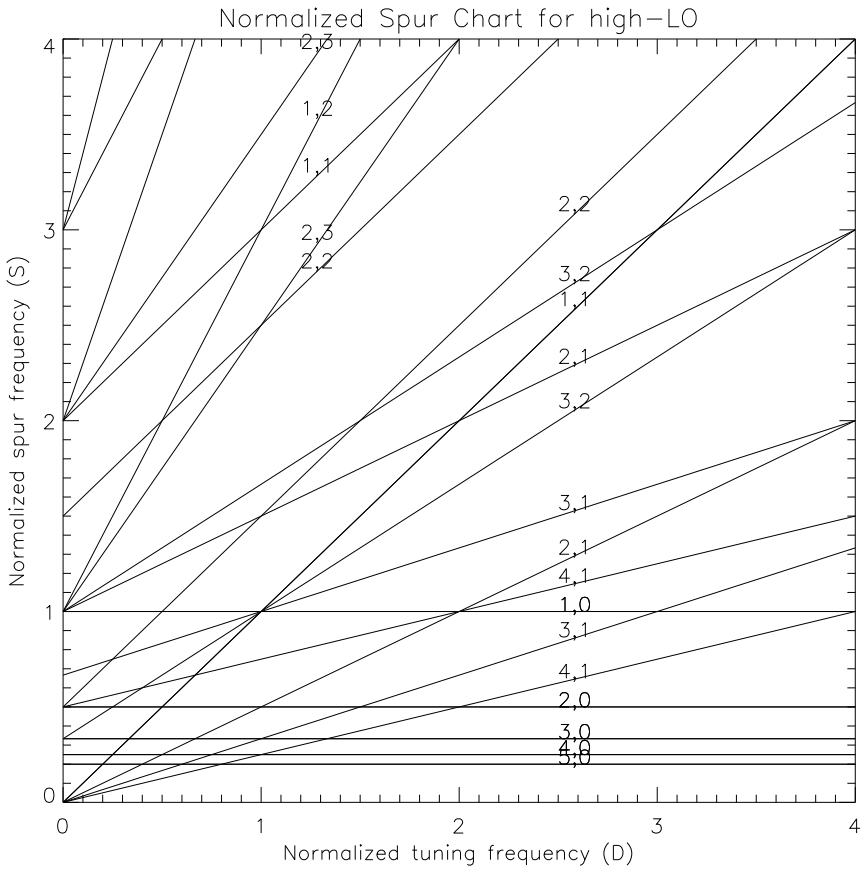


Figure 11.13: Universal spur chart for high LO. The numbers on each line (m, n) indicate the multiple of the RF and LO frequencies, respectively that are summed or differenced to produce an output at the IF. Only responses up through 5'th order $(m + n \leq 5)$ are shown. Normalized tuning frequencies greater than 1.0 correspond to downconversion and tuning frequencies less than 1.0 correspond to upconversion.

to tuning frequencies that are smaller than the IF, ($D = f_D/f_{IF} < 1$) and, hence, to up-conversion. The part of the plot where $D > 1$ corresponds to down-conversion.

Universal spur charts may also be produced for up-conversion with "low LO" and down-conversion with "low-LO". It turns out that the following normalized equations can be used for both of these cases. When $D < 1$ the result corresponds to up-conversion/low-LO and with $D > 1$ the result corresponds to down-conversion/low-LO:

$$\begin{aligned} S &= |\tfrac{1-n}{m} + \tfrac{n}{m}D| \\ S &= |\tfrac{1+n}{m} - \tfrac{n}{m}D| \end{aligned} \tag{11.30}$$

The universal spur chart for low-LO is shown in Figure 11.14.

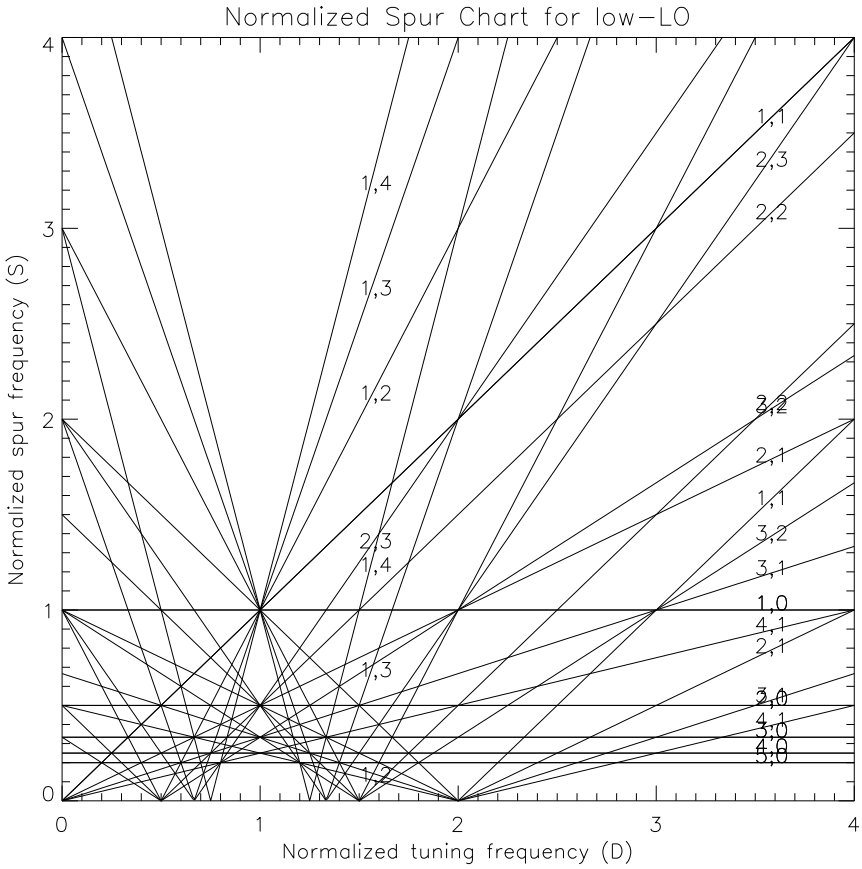


Figure 11.14: Universal spur chart for low LO. Only responses up through 5'th order are shown. The numbers on each line (m, n) indicate the multiple of the RF and LO frequencies, respectively that are summed or differenced to produce an output at the IF. Normalized tuning frequencies above 1.0 correspond to downconversion, and tuning frequencies below 1.0 correspond to upconversion.

11.4.1 Crossovers

Notice that both of the universal spur charts exhibit "crossovers", i.e. certain lines cross the m=1, n=1 line which corresponds to $S = D$. A crossover at some tuned frequency D means that a signal with frequency $S = D$ will interfere with itself! For example notice that a crossover occurs in Figure 11.13 when $S = D = 3$. In this case, the line that crosses the $m = 1$, $n = 1$ line at $S = D = 3$ has $m = 3$, $n = 2$. This means that the 3 harmonic of the desired signal at frequency $S = 3$ will mix with the second harmonic of the local oscillator to give an output at the IF. A numerical example will be provided in the next section. For now, it is sufficient to point out that a crossover frequency represents a potential "dead zone" in a receiver's tuning range, since it may be impossible to receive strong signals at this frequency.

11.4.2 Example - AM Broadcast band radio

Figure 11.15 shows a spur chart for a broadcast band AM radio employing high LO. For an IF of 455 kHz, the normalized lower and upper frequencies of the AM broadcast band are [540/455, 1700/455]=[1.187,3.736]. These limits are denoted in the figure by solid vertical and horizontal lines.

11.4.2.1 Radio tuned to receive a signal at 910 kHz:

Suppose the radio is tuned to receive a desired signal with carrier frequency 910 kHz (D=910/455=2.0). Consider possible interference from signals within the AM band resulting from spurious responses up through 5'th order (order = n+m). Examine the vertical dotted line drawn at the normalized tuned frequency D=2 and notice that for spur frequencies within the AM band there are 6 intersections with the dotted line (not counting the intersection with the desired n=m=1 line). This means there are 6 possible frequencies that could cause interference. If spurious response orders higher than 5 are considered, there would be an even larger number of potential spurious responses.

Suppose it is necessary to determine the exact frequencies of each of the potential spurs. It is not possible to obtain very accurate results using the chart, but the chart can be used to provide guidance for determining the numerical values of the spur frequencies. Notice that (m,n) for the 6 spurs are: (3,1), (3,2), (2,1), (3,2), (2,2), (2,2). Also notice that the (2,1) spur is a crossover spur. To determine the exact frequencies recall that spurs satisfy:

$$|mf_S \pm nf_{LO}| = f_{IF}$$

which corresponds to the three equations

$$mf_S + nf_{LO} = f_{IF} \tag{11.31}$$

$$mf_S - nf_{LO} = f_{IF} \tag{11.32}$$

$$-mf_S + nf_{LO} = f_{IF} \tag{11.33}$$

Since the AM radio uses high LO, equation 11.31 can never be satisfied and is not relevant here. The other two equations can be re-arranged to yield:

$$f_S = \frac{f_{IF} + nf_{LO}}{m} \tag{11.34}$$

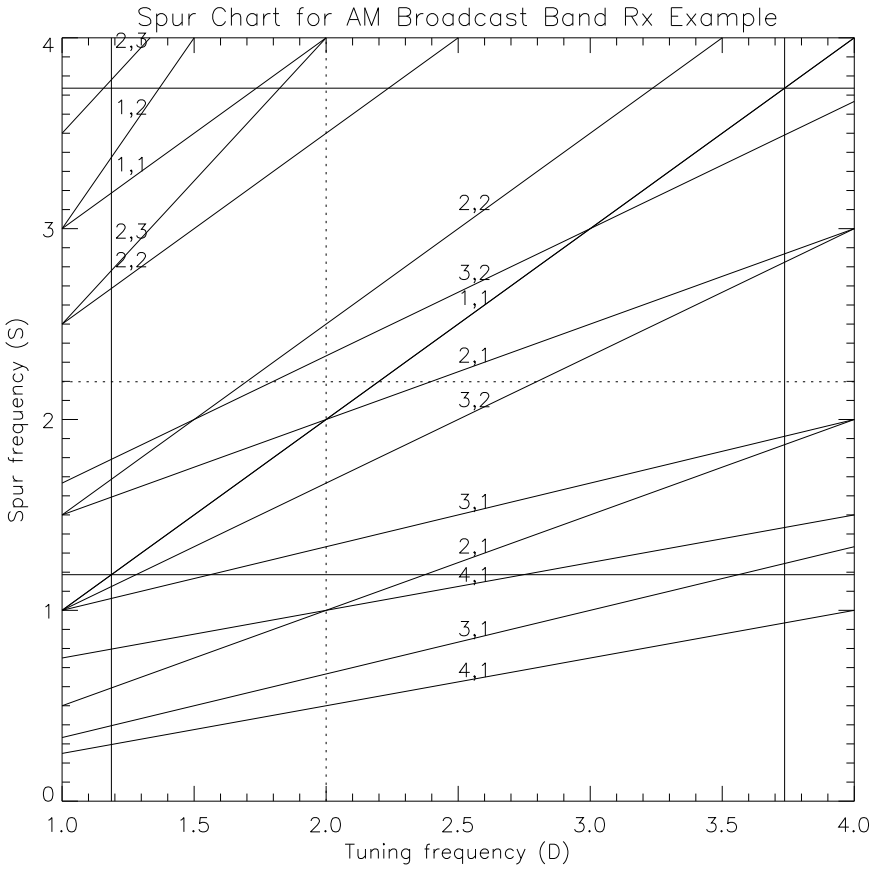


Figure 11.15: A portion of the universal spur chart for high-LO relevant to an AM Broadcast-band radio with 455 kHz IF and high LO.

$$f_S = \frac{nf_{LO} - f_{IF}}{m} \tag{11.35}$$

Now, use the fact that when the radio is tuned to 910 kHz, $f_{LO} = 910 + 455 = 1365$ kHz to find the exact frequency of each spur, i.e.:

$$(3,1) \Rightarrow f_S = \frac{455 + 1365}{3} = 606.67\,\text{kHz}$$

$$(3,2) \Rightarrow f_S = \frac{455 + 2(1365)}{3} = 1061.67\,\text{kHz}$$

$$(3,2) \Rightarrow f_S = \frac{2(1365) - 455}{3} = 758.33\,\text{kHz}$$

$$(2,2) \Rightarrow f_S = \frac{455 + 2(1365)}{2} = 1592.50\,\text{kHz}$$

$$(2,2) \Rightarrow f_S = \frac{2(1365) - 455}{2} = 1137.5\,\text{kHz}$$

$$(2,1) \Rightarrow f_S = \frac{455 + 1365}{2} = 910\,\text{kHz}$$

Notice that the (2,1) spur is a crossover spur, i.e. the spur frequency is equal to the desired frequency (910 kHz in this case). The interference results from the second harmonic of the 910 kHz signal (at 1820 kHz) mixing with the LO at 1365 kHz to produce an output from the mixer at 455 kHz. This self-interference would be expected to be most noticeable when the desired signal at 910 kHz is a very strong signal. It has the somewhat peculiar property that the stronger the desired signal, the more intense the interference will be, since the second harmonic will be generated most efficiently when the mixer is driven hard.

11.4.2.2 Strong signal at 1000 kHz

Next, suppose that a very strong signal is transmitting at 1000 kHz. As the receiver is tuned across the band, numerous spurious responses may occur. In this case, the strong signal at 1000 kHz (S=1000/455=2.1978) is represented by a horizontal dotted line. Notice that there are 4 possible tuned frequencies where spurs from the signal at 1000 kHz would be potentially observed. (The intersection between the horizontal dotted line and the (1,1) line is not counted, since that represents the case where the receiver is tuned to receive 1000 kHz.) The 4 spurious responses correspond to (2,2), (3,2), (2,1), (3,2). Equations 11.32 and 11.33 can be used along with the fact that $f_D = f_{LO} - f_{IF}$ to solve for the tuning dial setting at which the spurious responses would occur:

$$f_D = \frac{m}{n} f_S - f_{IF}(\frac{1}{n} + 1) \tag{11.36}$$

and

$$f_D = \frac{m}{n} f_S + f_{IF}(\frac{1}{n} - 1). \tag{11.37}$$

Using $f_S = 1000\,\text{kHz}$ and $f_{IF} = 455\,\text{kHz}$:

$$(2,2) \Rightarrow f_D = 1000 + 455(-\frac{1}{2}) = 772.5\,\text{kHz}$$

$$(3,2) \Rightarrow f_D = \frac{3}{2}1000 - 455(\frac{3}{2}) = 817.5\,\text{kHz}$$

$$(3,2) \Rightarrow f_D = \frac{3}{2}1000 + 455(-\frac{1}{2}) = 1272.5\,\text{kHz}$$

$$(2,1) \Rightarrow f_D = 2(1000) - 455(2) = 1090\,\text{kHz}$$

Thus, spurious responses would potentially be observed at the following dial settings: 772.5 kHz, 817.5 kHz, 1090 kHz, and 1272.5 kHz.

11.5 Homework Problems

1. Consider the mixer circuit shown in Figure 11.16:

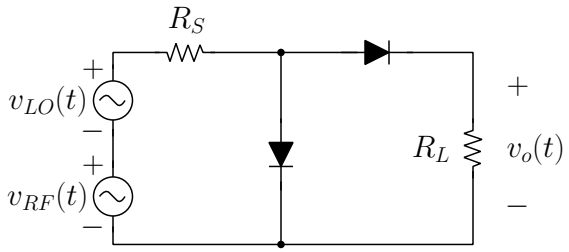


Figure 11.16: Mixer circuit

   (a) Assume that v_{LO} controls the state of the diodes at all times and that the diodes are ideal. Derive an expression for the output voltage. Express your result in terms of a switching function p(t), and give the Fourier cosine series for your switching function.

   (b) What frequency components will appear at the output of this mixer?

2. Consider the mixer circuit shown in Figure 11.17. The transformer is an ideal 3-winding transformer, which has the property that the voltages across all 3 windings are equal, with polarity corresponding to the dots shown next to each winding. The time-varying RF and LO voltages are $v_{RF}(t) = V_{RF}\cos\omega_{RF}t$, and $v_{LO}(t) = V_{LO}\cos\omega_{LO}t$. Assume that $v_{LO}(t)$ controls the state of the diodes at all times, and that the diodes are ideal switches.

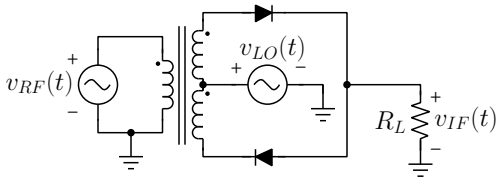


Figure 11.17: Mixer circuit

   (a) Derive an expression for the IF voltage, $v_{IF}(t)$. Express your result in terms of a switching function $p(t)$, and give the Fourier cosine series for your switching function.

   (b) List the frequencies of all components that will appear at the output of this mixer.

Chapter 12

Nonlinear Effects in 2-ports

Consider a time-invariant 2-port (Figure 12.1).

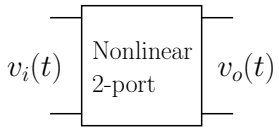


Figure 12.1: Nonlinear 2-port.

For sufficiently small input signals, the 2-port can often be modeled as a linear 2-port i.e.

$$v_o(t) = \int_{-\infty}^{\infty} h(\tau) v_i(t-\tau) d\tau \tag{12.1}$$

where $h(\tau)$ is the *impulse response* of the system. When the bandwidth of $v_i(t)$ is much smaller than the bandwidth of the 2-port, then the impulse response of the 2-port can be approximated as a delayed delta function, i.e. $h(t) \simeq k\delta(t-\tau_g)$ where k is the small-signal voltage gain and τ_g is the group-delay of the 2-port. In this case the output waveform will be a scaled and delayed version of the input signal, i.e.

$$v_o(t) \simeq k v_i(t-\tau_g).$$

Likewise, if $v_i(t) = s_1(t) + s_2(t)$, then

$$v_o(t) = k s_1(t-\tau_g) + k s_2(t-\tau_g) \tag{12.2}$$

i.e., superposition applies.

In general, for input signals that are not "small", the input-output relationship must be considered to be nonlinear. To illustrate some of the important characteristics of nonlinear 2-ports while keeping the analysis relatively simple, we will ignore the possibility of energy storage elements (such as inductances and capacitances) within the nonlinear 2-port. This means we can ignore frequency-dependent phase shifts and, therefore, delays. Such a 2-port will be frequency independent and is said to be *memoryless* because lack of energy storage means that the output can depend only on the present value of the input, and not on past

values. In many cases, this assumption can be justified by lumping energy storage elements into external, linear, networks. If the 2-port is time-invariant and memoryless, then the output signal will be related to the input signal by an nonlinear transfer characteristic, i.e. $v_o = f[v_i]$, where $f[\,]$ is a single-valued nonlinear function. Typically, the nonlinear input-output characteristic is approximated by a Taylor series in the vicinity of an operating point, and analysis proceeds based on the truncated Taylor series expansion. For input signals that are not too large, only a few terms of the Taylor series are sufficient. The Taylor series approach can be viewed as a special case of the more general *Volterra series* analysis which is applicable to nonlinear 2-ports with memory.

Denote the nonlinear input-output relationship for a memoryless 2-port as

$$v_o(t) = f[v_i(t)], \tag{12.3}$$

where $f[\,]$ is a single-valued, smooth (differentiable), nonlinear function. One consequence of the nonlinear input-output relationship for a memoryless 2-port is that superposition cannot be applied to determine the output in response to multiple inputs. In this case, if $v_i(t) = s_1(t) + s_2(t)$, then

$$\begin{aligned} v_o(t) &= f[s_1(t) + s_2(t)] && (12.4)\\ &\neq f[s_1(t)] + f[s_2(t)] && (12.5) \end{aligned}$$

i.e., superposition does not apply. As we shall see, when the input signal contains only one, or a few, sinusoidal components, the output signal may contain many more frequency components, generated because of the nonlinearity of the 2-port. When generated within receivers these distortion products can cause interference. In transmitters, nonlinearity in the power amplifier can cause significant broadening of the signal spectrum, possibly causing the output spectrum to bleed over into adjacent channels.

12.1 Power series model

For moderately large signals, a memoryless nonlinear input-output characteristic can be expanded in a Taylor series about some operating, or quiescent, point. Denote the total input and output signals by x_T and y_T, respectively so that $y_T = f[x_T]$. The input signal can be decomposed into quiescent, and time-varying components, i.e.

$$x_T = x_Q + x \tag{12.6}$$

where x_Q is the DC component of the input signal and x is a zero-mean, time-varying component. Similarly, denote the total output by y_T where

$$y_T = y_Q + y. \tag{12.7}$$

The first term, y_Q, is defined to be the output when $x = 0$. Then

$$\begin{aligned} y_T &= f[x_T] && (12.8)\\ &= f[x_Q] + x\left.\frac{df}{dx}\right|_{x=x_Q} + \frac{1}{2}x^2\left.\frac{d^2f}{dx^2}\right|_{x=x_Q} + \ldots + \frac{1}{n!}x^n\left.\frac{d^nf}{dx^n}\right|_{x=x_Q} + \ \ldots \end{aligned}$$

The part of the output that results from the time-varying part of the input is, therefore, of the form

$$y = k_1 x + k_2 x^2 + k_3 x^3 + \ldots \tag{12.9}$$

where

$$k_i = \left. \frac{d^i f}{dx^i} \right|_{x=x_q}$$

So-called *small-signal* models retain only the first term in equation 12.9, which represents a linear relationship between x and y. This term is dominant when x is small enough so that higher order terms are negligible. The higher order terms represent nonlinear distortion. Also, notice that even when x has zero mean, y may not have zero mean since terms which are even powers of the input signal will have non-zero means; for example, if the input is a cosine function $x = \cos \omega t$, the term $x^2 = \cos^2 \omega t = \frac{1}{2}(1 + \cos 2\omega t)$ is not zero-mean. Hence, the constant part of the output may contain terms contributed by the zero-mean input signal, x. This means that the DC bias point of an amplifier will shift when the amplifier is driven with a time-varying input signal.

To obtain a general understanding of nonlinear effects in 2-ports (without considering particular circuit configurations), we will assume that the input-output relationship can be accurately modeled by a power-series expansion

$$y = k_1 x + k_2 x^2 + k_3 x^3 + \ldots \tag{12.10}$$

where x represents the input excitation and y is the output quantity. Typically, x and y will represent input and output voltages or currents. For signals that are not too large it is sufficient to truncate the expansion to terms of third order and lower. Thus for the subsequent discussion we'll assume an input-output relationship consisting of a three-term power series:

$$y \simeq k_1 x + k_2 x^2 + k_3 x^3 \tag{12.11}$$

12.1.1 Specific Examples - BJT and FET nonlinearities

As an example, consider the input-output relationship for the bipolar junction transistor (BJT) in Figure 12.2.

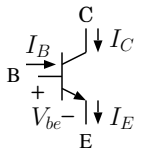


Figure 12.2: Bipolar junction transistor (BJT).

$$I_C = I_S e^{V_{be}/V_T} \tag{12.12}$$

and $V_T = \frac{kT}{q} \simeq 25\,\text{mV}$ at room temperature (290 K). Now suppose that the base-emitter voltage consists of a DC component (V_{DC}) and a time-varying signal component (v_{be}), i.e.,

$$V_{be} = V_{DC} + v_{be} \tag{12.13}$$

Then

$$I_C = I_S e^{V_{DC}/V_T} e^{v_{be}/V_T} \tag{12.14}$$

If $v_{be}/V_T < 1$, we can expand e^{v_{be}/V_T}:

$$I_C = I_S e^{V_{DC}/V_T} \left[1 + \frac{v_{be}}{V_T} + \frac{1}{2}\left(\frac{v_{be}}{V_T}\right)^2 + \frac{1}{6}\left(\frac{v_{be}}{V_T}\right)^3 + \ldots \right] \tag{12.15}$$

The first term represents the DC component of collector current when $v_{be} = 0$. Denote this *quiescent* current by I_{CQ}, i.e.:

$$I_{CQ} = I_S e^{V_{DC}/V_T}.$$

Denote the fluctuating component of the collector current by i_c, then a three-term power-series approximation for the nonlinear relationship between i_C and v_{be} is:

$$i_C = \frac{I_{CQ}}{V_T} v_{be} + \frac{I_{CQ}}{2V_T^2} v_{be}^2 + \frac{I_{CQ}}{6V_T^3} v_{be}^3$$

Hence, for the BJT:

$$\begin{aligned} k_1 &= \tfrac{I_{CQ}}{V_T} = g_m \\ k_2 &= \tfrac{I_{CQ}}{2V_T^2} \\ k_3 &= \tfrac{I_{CQ}}{6V_T^3} \end{aligned} \tag{12.16}$$

The first term is the familiar small-signal approximation; the time-varying part of the collector current is equal to the transconductance, g_m, times the time-varying part of the base-emitter voltage. In this chapter we will be concerned with the implications of the higher order terms in the input-output characteristic.

It is interesting to compare the BJT input-output characteristic with that for an ideal field effect transistor (FET) as in Figure 12.3.

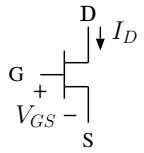


Figure 12.3: Field-effect transistor (FET).

$$I_D = I_{DSS}\left(1 - \frac{V_{GS}}{V_P}\right)^2 \tag{12.17}$$

$$= I_{DSS}\left(1 - \frac{2}{V_P} V_{GS} + \frac{1}{V_P^2} V_{GS}^2\right) \tag{12.18}$$

There are no third-order or higher terms in this idealized *square-law* FET characteristic. In a more realistic model for the FET such terms would be present.

12.2 Single-tone Input

Consider a nonlinear amplifier modeled with a three term power-series:

$$v_o = k_1 v_i + k_2 v_i^2 + k_3 v_i^3$$

Suppose the input signal consists of a single tone

$$v_i(t) = a_1 \cos \omega_1 t \tag{12.19}$$

Then

$$v_o(t) = k_1 a_1 \cos \omega_1 t \; + k_2 a_1^2 \cos^2 \omega_1 t + k_3 a_1^3 A \cos^3 \omega_1 t \tag{12.20}$$

Using trigonometric identities

$$\begin{aligned} \cos^2 \omega t &= \frac{1}{2}(1 + \cos 2\omega t) \\ \cos^3 \omega t &= \frac{1}{4}(3 \cos \omega t + \cos 3\omega t) \end{aligned} \tag{12.21}$$

The output is written

$$\begin{aligned} v_o(t) &= \tfrac{1}{2} k_2 a_1^2 + (k_1 a_1 + \tfrac{3}{4} k_3 a_1^3) \cos \omega_1 t + \\ &\quad \tfrac{1}{2} k_2 a_1^2 \cos 2\omega_1 t + \tfrac{1}{4} k_3 a_1^3 \cos 3\omega_1 t \end{aligned} \tag{12.22}$$

and the input and output spectra are illustrated in Figure 12.4. Since the amplitude of the second and third harmonics is proportional to a_1^2 and a_1^3, respectively, this analysis predicts that for every 1 dB increase of the input signal, the second and third harmonics will increase by 2 dB and 3 dB, respectively.

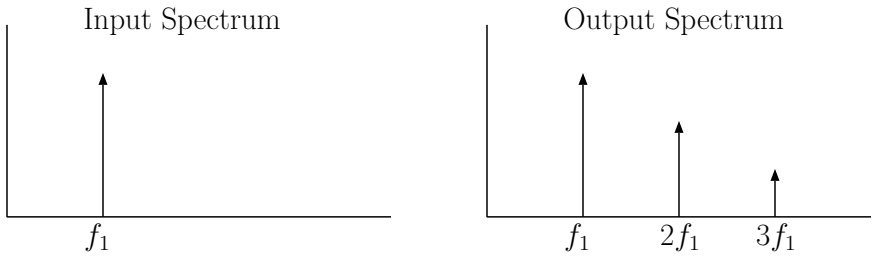


Figure 12.4: Input and output spectra for single input signal.

12.2.1 Gain Compression

Now consider the output component at the fundamental frequency:

$$\begin{aligned} v_o'(t) &= (k_1 a_1 + \tfrac{3}{4} k_3 a_1^3) \cos \omega_1 t \\ &= k_1 (1 + \tfrac{3}{4} \tfrac{k_3}{k_1} a_1^2) a_1 \cos \omega_1 t \end{aligned} \tag{12.23}$$

The effective voltage gain of the amplifier is

$$A_v = k_1(1 + \frac{3}{4}a_1^2\frac{k_3}{k_1}) \tag{12.24}$$

The term in parenthesis arises because the amplifier exhibits a third-order nonlinearity and its magnitude depends on the input signal amplitude. Depending on the sign of k_3, as the input signal amplitude increases the effective gain of the amplifier will either increase (*expansive nonlinearity*) or decrease (*compressive nonlinearity*). If k_1 and k_3 have the same sign, then the gain will increase with increasing input amplitude, and if k_1 and k_3 have opposite signs, the gain decreases with increasing input amplitude. Our analysis of the BJT in section 12.1.1 showed that k_1 and k_3 have the same sign, and hence the nonlinearity is *expansive*. The analysis in section 12.1.1 is based on the assumption that the DC component of V_{be} is held constant. In practical amplifier circuits, the transistor will be biased such that the DC component of the collector current is held more-or-less constant (constant-current bias), and the DC component of V_{be} is allowed to adjust in order to maintain constant DC collector current. Analysis based on this constraint will show that such an amplifier exhibits *compressive* nonlinearity. A detailed analysis of this type is carried out in Appendix A, where the *large signal transconductance* is derived and shown to decrease with increasing base-emitter voltage swing. For now, we shall simply state that in most cases k_1 and k_3 will have opposite signs, so that the nonlinearity is compressive. For compressive nonlinearity, the effective gain can be written as

$$k_1\left(1 - \frac{3}{4}a_1^2\left|\frac{k_3}{k_1}\right|\right)$$

The reduction in gain with increasing input signal amplitude caused by compressive nonlinearity is called *gain compression*. The plot in Figure 12.5 shows the output power in the fundamental component as a function of the input power for an amplifier that exhibits gain compression.

The gain compression of a 2-port is often characterized by the input power level that causes the gain to be decreased by 1dB. This parameter is denoted by P_{1dB} in Figure 12.5.

12.3 Two-tone Input

Now suppose that the input signal includes two tones:

$$v_i(t) = a_1\cos\omega_1 t + a_2\cos\,\omega_2 t \tag{12.25}$$

Then

$$\begin{aligned} v_o(t) \quad = \quad & k_1[a_1\cos\omega_1 t + a_2\cos\omega_2 t] \\ & + k_2[a_1\cos\omega_1 t + a_2\cos\omega_2 t]^2 \\ & + k_3[a_1\cos\omega_1 t + a_2\cos\omega_2 t]^3 \end{aligned} \tag{12.26}$$

Using trigonometric identities the output can be re-written as

$$v_o(t) = \text{first order terms} + \text{second order terms} + \text{third order terms}$$

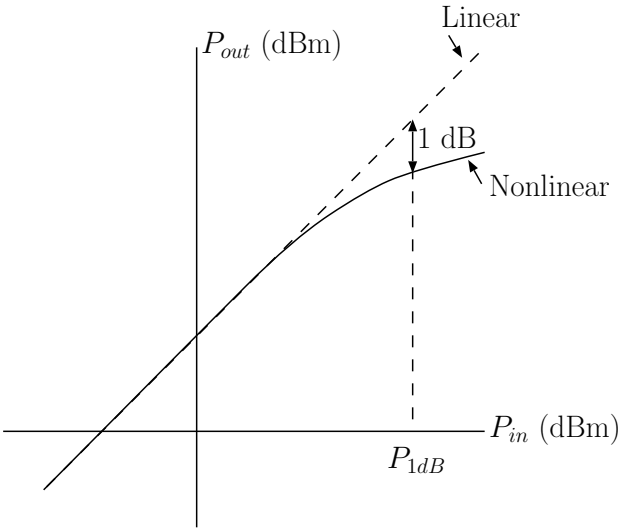


Figure 12.5: Output power in fundamental component as a function of input power.

where

$$\text{first order terms} = k_1[a_1 \cos\omega_1 t + a_2 \cos\omega_2 t]$$

$$\begin{aligned}\text{second order terms} &= k_2[\tfrac{1}{2}(a_1^2 + a_2^2) + \tfrac{1}{2}a_1^2 \cos 2\omega_1 t + \tfrac{1}{2}a_2^2 \cos 2\omega_2 t \\ &\quad + a_1 a_2 \cos(\omega_1 + \omega_2)t + a_1 a_2 \cos(\omega_1 - \omega_2)t]\end{aligned}$$

$$\begin{aligned}\text{third order terms} &= k_3[(\tfrac{3}{4}a_1^3 + \tfrac{3}{2}a_1 a_2^2)\cos\omega_1 t + (\tfrac{3}{4}a_2^3 + \tfrac{3}{2}a_1^2 a_2)\cos\omega_2 t \\ &\quad + \tfrac{1}{4}a_1^3 \cos 3\omega_1 t + \tfrac{1}{4}a_2^3 \cos 3\omega_2 t \\ &\quad + \tfrac{3}{4}a_1 a_2^2(\cos(2\omega_2 - \omega_1)t + \cos(2\omega_2 + \omega_1)t) \\ &\quad + \tfrac{3}{4}a_1^2 a_2(\cos(2\omega_1 - \omega_2)t + \cos(2\omega_1 + \omega_2)t)]\end{aligned}$$

The input and output spectra are illustrated in Figure 12.6.

In narrow-band systems most of these frequency components will be removed by filters downstream from the nonlinear 2-port, but the so-called in-band third-order terms ($2f_2 - f_1$, $2f_1 - f_2$) cannot be ignored since they will always have frequencies close to that of f_1 and f_2. In wide-band systems, all of the terms can potentially be significant at the output. In receivers that use a low-IF, the second-order term at the difference frequency $f_2 - f_1$ may be

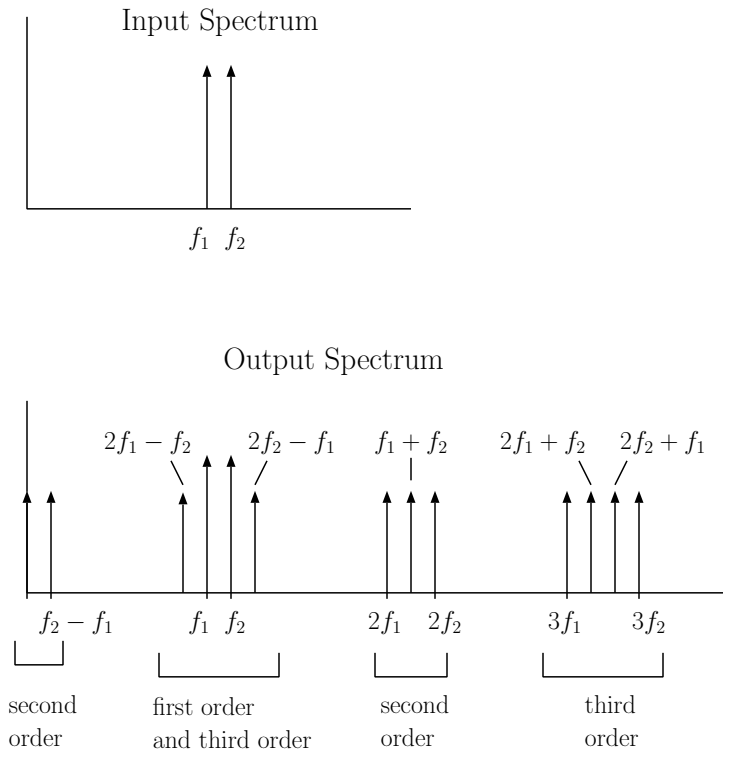


Figure 12.6: Input (top) and output (bottom) spectra for two input signals.

close to the IF, and is of special concern. For now we will consider only the in-band terms:

$$\begin{aligned} v_o'(t) &= \left\{k_1 a_1 + k_3\left(\frac{3}{4}a_1^3 + \frac{3}{2}a_1 a_2^2\right)\right\}\cos\omega_1 t \\ &\quad + k_3\frac{3}{4}a_1 a_2^2\cos(2\omega_2 - \omega_1)t \\ &\quad + k_3\frac{3}{4}a_1^2 a_2\cos(2\omega_1 - \omega_2)t \\ &+ \left\{k_1 a_2 + k_3\left(\frac{3}{4}a_2^3 + \frac{3}{2}a_1^2 a_2\right)\right\}\cos\omega_2 t \end{aligned} \tag{12.27}$$

There are a number of phenomena that can be illustrated using this result.

12.3.1 Desensitization and Blocking

Suppose $a_1\cos\omega_1 t$ is a relatively weak desired signal and $a_2\cos\omega_2 t$ is a strong signal at a nearby frequency that is not of interest. To model this situation we assume that $a_2 \gg a_1$. Taking into account that $a_2 \gg a_1$, and assuming that the amplifier exhibits compressive nonlinearity, the amplitude of the desired $\cos\omega_1 t$ term is, approximately:

$$\approx k_1 a_1 + k_3\frac{3}{2}a_1 a_2^2 \tag{12.28}$$

$$= k_1 a_1\left(1 - \left|\frac{k_3}{k_1}\right|\frac{3}{2}a_2^2\right) \tag{12.29}$$

Notice that the amplitude of the desired term will be reduced as the amplitude a_2 increases. This means that the presence of $a_2\cos\omega_2 t$ causes gain compression for the desired signal, even though the desired signal may be too weak to cause gain compression by itself. Thus, an undesired strong signal can desensitize a receiver to the presence of weak signals by compressing the gain. This phenomenon is called *densensitization*, or *blocking*.

12.3.2 Cross modulation

Another undesirable effect can result if both signals are modulated. For example, suppose that the signals are amplitude modulated, i.e., $a_1 \Rightarrow a_1(1 + m_1(t))$ and $a_2 \Rightarrow a_2(1 + m_2(t))$ and that $a_2 \gg a_1$. Then the envelope of the desired signal will be given by

$$= k_1 a_1(1 + m_1(t)) + k_3\frac{3}{2}a_1(1 + m_1(t))a_2^2(1 + m_2(t))^2 \tag{12.30}$$

Notice that the signal at frequency ω_1 now contains amplitude modulation from signal at frequency ω_2. The modulation from the undesired strong signal at ω_2 has been transferred to the desired signal at ω_1. This phenomenon is called *cross modulation*. If the signals are angle modulated, the same mechanism will cause the angle modulation on signal 1 to contain a term from the angle modulation on signal 2.

12.3.3 More than two tones and nonlinear terms with order higher than 3

So far we have considered only two input signals. If there are more than two input tones, many more intermodulation products will be generated. In general, for M input tones ($\{f_i\},\ i = 1,\ 2,\ \ldots\ M$) an n'th order non-linearity will generate all positive frequencies given by the magnitude of the sum of n terms of the form $\pm f_i,\ i \in \{1, 2, \ldots M\}$. For example, if there are 3 input signals at frequencies $f_1,\ f_2,\ f_3$ the third-order nonlinearity will produce third-order products at $|\pm f_i \pm f_j \pm f_k|$, where $i,\ j,\ k\ \in \{1,\ 2,\ 3\}$. When the three input frequencies are closely spaced, the frequencies resulting when two of the three terms have the same sign will be close to the input frequencies and, hence, are called in-band intermodulation (IM) products. The in-band third order IM products when there are three input tones are listed below:

$$2f_1 - f_2,\ \ 2f_2 - f_1$$

$$2f_3 - f_1,\ \ 2f_1 - f_3$$

$$2f_2 - f_3,\ \ 2f_3 - f_2$$

$$f_1 + f_2 - f_3,\ \ f_1 + f_3 - f_2,\ \ f_2 + f_3 - f_1$$

$$f_1,\ \ f_2,\ \ f_3$$

With three input tones, a 4'th order nonlinearity will generate frequencies of the form $|\pm f_i \pm f_j \pm f_k \pm f_l|$ where $i,\ j,\ k,\ l\ \in \{1,\ 2,\ 3\}$, etc. Note that the even-order nonlinearities will not produce in-band products. Only the odd-order nonlinearities are responsible for in-band IM products. For example, with two input tones, a 5'th order nonlinearity would produce output components at $|f_1 + f_1 + f_1 - f_2 - f_2| = 3f_1 - 2f_2$, which is an in-band IM product.

12.4 Quantitative Characterization of IM Distortion

When multiple input signals are present, nonlinearity causes complex effects. For quantitative comparisons between 2-ports, an idealized situation, or special case, is often used to provide a common basis for comparisons. The usual conditions for this test are to apply two input signals with equal amplitudes; this is called a "two-tone" intermodulation distortion (IMD) test. The input signals can be written as

$$v_i(t) = a(\cos\omega_1 t + \cos\omega_2 t) \tag{12.31}$$

It is assumed that the amplitude of the two signals is small enough so that gain compression doesn't occur. Then the in-band products at the output of the 2-port are

$$\begin{aligned} v_o'(t) \quad &= \quad k_1 a(\cos\omega_1 t + \cos\omega_2 t) \qquad (12.32) \\ &\quad + k_3 \frac{3}{4} a^3(\cos(2\omega_2 - \omega_1)t + \cos(2\omega_1 - \omega_2)t) \end{aligned}$$

Note that the coefficient of the fundamental terms is simply k_1 which is the result of ignoring gain compression. Next define the input power of each tone, P_{in}. For simplicity, we shall assume that the input impedance of the 2-port is resistive, and equal to the value R_{in}. Then

$$P_{in} = \frac{1}{2}a^2 R_{in}^{-1} \tag{12.33}$$

Assuming that the load impedance is resistive, and denoted by R_L, the output power of each of the desired (fundamental) components is then

$$P_d = \frac{1}{2}k_1^2 a^2 R_L^{-1}$$

The output power of the in-band third-order intermodulation products is

$$P_{IM} = \frac{1}{2}k_3^2(\frac{3}{4})^2 a^6 R_L^{-1}$$

Notice that P_d will be proportional to P_{in}, whereas P_{IM} is proportional to P_{in}^3. Figure 12.7 shows P_d and P_{IM} plotted against P_{in} on a log-log plot (dBm versus dBm). At the lower input power levels, P_{IM} increases with a slope of 3 compared to P_d which has a slope of 1. At higher input power, both components will eventually begin to saturate, as shown by the deviations of the curves from linearity. It should be noted that our simple 3-term power series model does not predict saturation of the IM products. It is necessary to include higher order terms to predict this behavior.

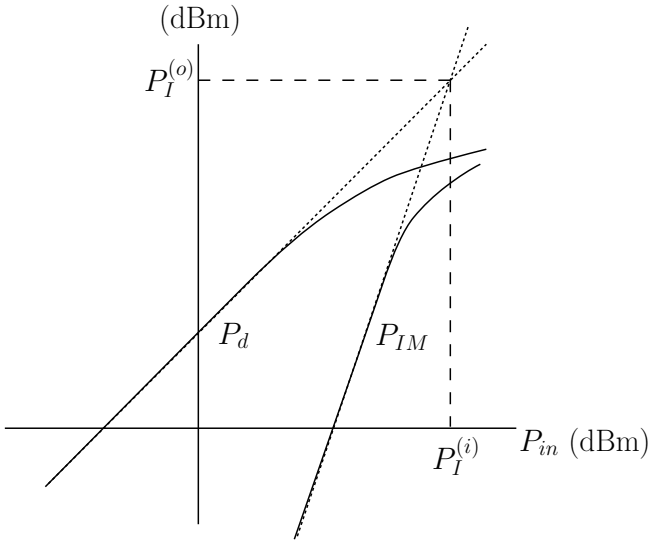


Figure 12.7: Output power (per tone) in the desired (P_d) and the IM (P_{IM}) components verses input power per tone. The third order intercept power, $P_I^{(i)}$, is determined by extrapolation from measurements at a small input power, where gain compression is negligible.

The fact that P_{IM} increases faster than P_d means that if gain compression did not eventually occur, the two output curves would ultimately intersect, as shown in Figure 12.7.

The input power level at which the linearly extrapolated curves intersect is called the input two-tone third-order intercept level and is denoted by $P_I^{(i)}$, or IIP3. The corresponding output power level is denoted by $P_I^{(o)}$, or OIP3.

It is important to note that the "intercept point" is fictitious, since we ignored gain compression. In actuality, both the P_{IM} and P_d curves would saturate at some finite value. This is ignored when computing the third-order intercept level. Since we know the slopes of the ideal P_{IM} and P_d curves, it is a simple matter to measure P_{IM} and P_d at a relatively low input power level where gain compression is not important. Then the curves can be extrapolated to the (fictitious) intercept point.

The intercept level is commonly used as a figure of merit for comparing the relative quality of amplifiers or systems. If it is known, then a system designer can determine P_{IM}/P_d for any two-tone input level where gain compression can be ignored. Two-ports are often operated at levels where gain compression is not important but where intermodulation distortion is still of concern.

Equation 12.34 summarizes the relationship between the input power level and the so-called intermodulation ratio, or IMR

$$IMR = \frac{P_{IM}}{P_d} = \left(\frac{P_{in}}{P_I^{(i)}} \right)^2 \tag{12.34}$$

Equation 12.34 can be used to determine P_{IM}/P_d given P_{in}, if $P_I^{(i)}$ is known. This is useful for system design applications. On the other hand, given measurements of P_d, P_{IM} for a known P_{in}, Equation 12.34 can also be used to find $P_I^{(i)}$.

12.4.1 Example - Calculating IMR

An amplifier has a two-tone third-order input intercept (IIP3) of 20 dBm. What is the intermodulation ratio (in dB) for a two-tone input level of 0 dBm?

The intermodulation ratio is calculated from:

$$IMR = \frac{P_{IM}}{P_d} = \frac{P_{in}^2}{P_I^{(i)2}} \tag{12.35}$$

or, in dB

$$\begin{aligned} IMR(\text{dB}) &= 2\,P_{in}\,(\text{dBm}) - 2\,P_I^{(i)}\,(\text{dBm}) && (12.36) \\ &= 2\,(0) - 2\,(20) \\ &= -40\,\text{dB} && (12.37) \end{aligned}$$

This means that the IM products (in-band) are 40 dB below the desired signals at the output, i.e., the output spectrum would look like Figure 12.8.

The next example shows how laboratory measurements of intermodulation products can be used to calculates the intercept point of a 2-port.

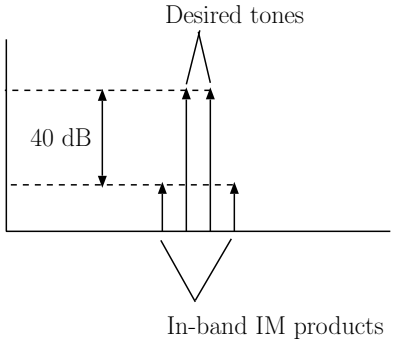


Figure 12.8: Output spectrum showing desired and IM products.

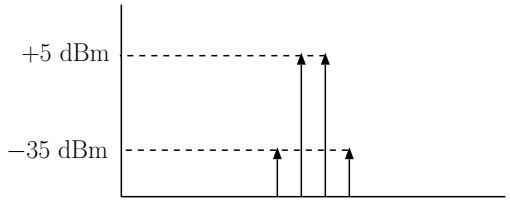


Figure 12.9: Output spectrum with 2 input signals, $P_{in} = -10\,\text{dBm}$.

12.4.2 Example - Calculating IIP3 ($P_I^{(i)}$).

Figure 12.9 shows the spectrum measured at the output of a 2-port when the input consists of 2 signals with $P_{in} = -10$ dBm. The power gain and input intercept of the 2-port can be found if it is known that gain compression can be neglected when interpreting the measurements. This can be verified by increasing the power of both input tones by 1 dB, and verifying that the desired tones at the output also increase by 1 dB, and that the in-band 3rd order intermodulation products each increase by 3 dB. If so, then

$$\begin{aligned} G &= +5\,\text{dBm} - (-10\,\text{dBm}) \\ &= 15\,\text{dB} \end{aligned} \tag{12.38}$$

The input intercept can be found from

$$P_{IM} - P_d = 2P_{in} - 2P_I^{(i)} \tag{12.39}$$

where all quantities are expressed in dBm. Solving for the input intercept:

$$\begin{aligned} P_I^{(i)} &= P_{in} + \frac{1}{2}(P_d - P_{IM}) \\ &= -10 + \frac{1}{2}(5 - (-35)) \\ &= +10\ \text{dBm} \end{aligned} \tag{12.40}$$

When performing a measurement like the one illustrated in this example, a typical setup will use a passive signal combiner to combine the outputs from two signal generators in order to construct the two-tone signal. It is important to take care to ensure that the combiner provides significant isolation between the two signal generators, so that intermodulation products are not generated within the signal generator output stages. It is also important to verify that intermodulation products generated within the spectrum analyzer can be neglected.

12.5 Dynamic Range of a Receiving System

The nonlinear effects that have been described in this chapter will determine the largest input signal level that a receiving system can handle without serious degradation of performance. The noise figure of the receiver, on the other hand, determines the smallest usable input signal level. Together the nonlinear effects and the noise floor determine the dynamic range of a receiver. Consider the receiving system in Figure 12.10.

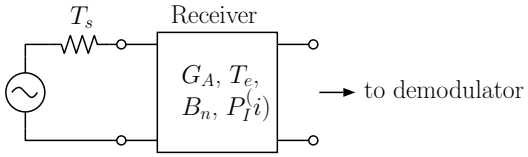


Figure 12.10: Receiving system.

The dynamic range (DR) is defined as follows:

$$DR = \frac{\text{maximum useable input signal power level}}{\text{minimum useable input signal power level}} \tag{12.41}$$

and is usually expressed in dB, i.e.,

$$DR = 10 \log \frac{P_{max}}{P_{min}} \tag{12.42}$$

The numerator P_{max} is the largest usable input signal power level. There are various definitions for P_{max} that depend on what nonlinear effect is being considered by the designer and what level of degradation is considered unsatisfactory. A commonly employed definition for the dynamic range is the "spurious free" dynamic range where the powers P_{max} and P_{min} are defined as follows:

P_{min} $\Rightarrow$ input power for specified $SNR_{o,min}$ at system output (same as MDS)
P_{max} $\Rightarrow$ two-tone input power (per tone) at which SNR of in-band 3rd-order IM products is equal to $SNR_{o,min}$ at system output.

The minimum usable input signal power is equal to the MDS, which has been defined previously:

$$P_{min} \equiv MDS = k(T_S + T_e)B_n SNR_{o,min}.$$

The maximum usable input signal power, P_{max}, is defined based on the two-tone input test concept. Suppose that the input consists of two equal amplitude tones. Then P_{max} is the input level that would cause the in-band third-order products to be detectable at the output of the receiver. In this context "detectable" means that the signal-to-noise ratio of the third-order products would be equal to the minimum SNR required for detection, $SNR_{o,min}$.

Assume that P_{min} has already been computed. The output signal power that results from an input power P_{min} is just $G_A P_{min}$. This output power represents the *detection threshold*. Now we ask what two-tone input power is required to cause the in-band third order products to have an output power equal to the detection threshold. The required input power is, by definition, P_{max}. When the input power, per tone, is P_{max}, the output power in the intermodulation products is $G_A P_{min}$. Using equation 12.34:

$$\frac{G_A P_{min}}{P_d} = (\frac{P_{max}}{P_I^{(i)}})^2.$$

Use the fact that $P_d = G_A P_{max}$:

$$\frac{P_{min}}{P_{max}} = (\frac{P_{max}}{P_I^{(i)}})^2.$$

Hence

$$P_{max} = P_{min}^{1/3} \left(P_I^{(i)}\right)^{2/3} \tag{12.43}$$

Thus, the dynamic range is

$$\begin{aligned} DR &= \frac{P_{min}^{1/3} \left(P_I^{(i)}\right)^{2/3}}{P_{min}} \\ &= \left(\frac{P_I^{(i)}}{P_{min}}\right)^{2/3} \end{aligned} \tag{12.44}$$

or, in dB

$$DR = \frac{2}{3}\left[P_I^{(i)} - P_{min}\right] \tag{12.45}$$

where $P_I^{(i)}$ and P_{min} are expressed in dBm. Recall, P_{min} is just the noise-floor of the receiving system and that it depends on the signal-to-noise ratio that is required for the particular type of signal and demodulator used in the system.

12.6 Intercept Point of a Cascade

Suppose that an amplifier with known input intercept level is cascaded with another amplifier as in Figure 12.11.

To simplify the analysis, note that the signal level to stage 2 is higher than to stage 1, if stage 1 has gain. So the nonlinear distortion in amplifier 2 is likely to be more important. We will ignore distortion generated in the first stage, which is equivalent to assuming that stage 1 is a linear preamplifier (equivalent to letting $P_{I1}^{(i)} \to \infty$).

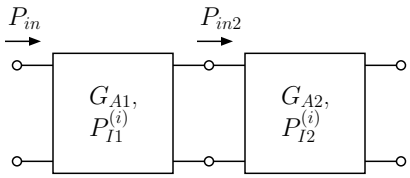


Figure 12.11: Cascaded amplifiers

Then, for the second stage alone:

$$\frac{P_{IM}}{P_d} = (\frac{P_{in2}}{P_{I2}^{(i)}})^2 \tag{12.46}$$

For the cascade, denote $P_I^{(i)}$ as the input intercept (as yet unknown), then

$$\frac{P_{IM}}{P_d} = (\frac{P_{in}}{P_I^{(i)}})^2 \tag{12.47}$$

Rewriting Equation 12.46 using $P_{in2} = G_{A1}\, P_{in}$

$$\begin{aligned} \frac{P_{IM}}{P_d} &= \frac{(G_{A1}\, P_{in})^2}{P_{I2}^2} \\ &= \frac{P_{in}^2}{(P_{I2}/G_{A1})^2} \end{aligned} \tag{12.48}$$

Comparing Equations 12.48 and 12.47 we conclude

$$P_I^{(i)} = \frac{P_{I2}^{(i)}}{G_{A1}} \tag{12.49}$$

The addition of stage 1 reduces the intercept point by a factor equal to the gain of the first stage. For example, if $P_{I2}^{(i)} = +6\,\text{dBm}$, $G_{A1} = 10\,\text{dB}$, then the cascade has input intercept $P_I = -4\,\text{dBm}$. In the next section, we will show that adding gain in front of an existing receiver will reduce the system DR.

12.6.1 The effect of adding a preamp to a receiver

Suppose that a receiver has effective input temperature T_{er}, noise bandwidth B_n, and input intercept $P_I^{(i)}$. The dynamic range of the receiver is then

$$DR_{receiver} = [\frac{P_I^{(i)}}{k(T_s\, +\, T_{er})B_n SNR_{o,min}}]^{2/3} \tag{12.50}$$

Suppose that a preamp with effective input temperature T_{ep} and available gain G_A is added before the receiver. Assuming that distortion generated in the preamplifier is negligible, the new dynamic range is

$$DR_{receiver\ with\ preamp} = [\frac{P_I^{(i)}/G_A}{k(T_s + T_{ep} + T_{er}/G_A)B_n SNR_{o,min}}]^{2/3} \tag{12.51}$$

Consider the ratio$DR_{with\ preamp}/DR_{receiver}$:

$$\frac{DR_{receiver\ with\ preamp}}{DR_{receiver}} = [\frac{T_s + T_{er}}{G_A(T_s + T_{ep}) + T_{er}}]^{2/3}. \tag{12.52}$$

Now, we ask whether this ratio can exceed unity, i.e. whether adding a preamp can increase the dynamic range of a receiving system. This requires

$$T_s + T_{er} > G_A(T_s + T_{ep}) + T_{er} \tag{12.53}$$

Subtracting T_s from both sides of Equation 12.53 results in the following inequality which must be satisfied, if the addition of a preamp is to increase the DR of a receiver:

$$T_{ep} < T_s(1 - G_A)/G_A \tag{12.54}$$

This inequality can never be satisfied if the preamp has available gain > 1, since the RHS of Equation 12.54 will be negative. The conclusion is that adding a preamp to an existing receiver will always decrease the DR of the receiver (although it may decrease the noise floor and therefore improve the sensitivity).

If the available gain of the added stage is less than 1, then the inequality can be satisfied. For example, suppose that the added stage is a passive attenuator. Then T_{ep} must be replaced with $T_{att}(\frac{1}{G_A} - 1)$ and the inequality is satisfied if

$$T_{att} < T_s.$$

Therefore, adding an attenuator in front of an existing receiver will increase the dynamic range of the system if the physical temperature of the attenuator is smaller than the source temperature (usually the antenna temperature). Adding an attenuator in front of an existing receiver will decrease the sensitivity (increase the MDS) of the receiver, however.

The intercept point of a receiver is usually set by the input intercept of the first mixer. The reason for this is as follows – intermodulation distortion is most likely to be generated in the stages ahead of the relatively narrow band IF amplifiers, i.e., the preamplifier and mixer stages. Since the input signal level at the mixer is generally larger than for any preceeding stage, the intercept point of the mixer will usually determine the largest usable input signal level. After the mixer, the signals are filtered by the IF filters, and generation of further intermodulation components in the IF amplifiers is less likely since, presumably, the desired signal component has been "picked out" from among other components. So the mixer is the most likely source of intermodulation distortion products (in a well designed receiver). Generation of intermodulation products in the mixer will be minimized if the gain in front of the mixer is as small as possible.

This leads us to an important principle for receiver design: The smallest possible amount of gain (preamp gain) should be used in front of the mixer. The minimum preamp gain will be determined by the preamp noise temperature and/or the antenna temperature together with the required MDS. Increasing the preamplifier gain above that required to set the noise floor to an acceptable level will only increase the input signal levels at the mixer input and degrade the large-signal handling capability of the receiver without providing useful improvement in sensitivity.

12.7 References

1. Carson, Ralph S., *Radio Concepts: Analog*, John Wiley & Sons, New York, 1990.

2. Clarke, Kenneth K. and Donald T. Hess, *Communication Circuits: Analysis and Design*, Addison-Wesley, 1978.

3. Maas, Stephen A., *Nonlinear Microwave Circuits*, Artech House, 1988.

4. Smith, Jack, *Modern Communications Circuits*, McGraw Hill, 1986.

12.8 Homework Problems

1. A nonlinear amplifier has an input-output voltage characteristic

$$v_o(t) = k_1 v_i(t) + k_2 v_i^2(t) \tag{12.55}$$

Suppose that three signals with equal amplitudes are input to the amplifier. The signal frequencies are

$$\begin{aligned} f_1 &= 0.6\,MHz \\ f_2 &= 1.3\,MHz \\ f_3 &= 1.5\,MHz \end{aligned} \tag{12.56}$$

List the frequencies of all components that will appear at the output of the amplifier.

2. Consider an AM broadcast band receiver that is designed to cover the frequency range 540 to 1700 kHz. The single-conversion superhet receiver uses an IF of 455 kHz and High LO. The receiver uses an RF amplifier that has the following nonlinear input-output characteristic:

$$v_o = k_1 v_{RF} + k_2 v_{RF}^2 + k_3 v_{RF}^3 \tag{12.57}$$

Suppose two signals at frequencies $f_1 = 720$ kHz and $f_2 = 780$ kHz are input to the receiver. Assume that both signals pass through the preselector. You may assume that the receiver stages that follow the RF amplifier are linear and that the mixer can be modeled as an ideal multiplier. List the frequencies of all possible output signals from the RF amplifier that could be detected with the receiver. Also, for each signal that can be detected, list the receiver tuning dial setting at which the signal will be detected.

3. A nonlinear 2-port's input-output characteristic can be modeled as

$$v_o = 12 v_i - v_i^3$$

where v_i and v_o represent the instantaneous values of the time-varying component of the input and output voltage, respectively.

   (a) Suppose that this 2-port is used in a system with source and load impedance of 50 Ω. Assume that the 2-port is conjugately matched at the input and output in this system. Find the input power level that causes 1dB of gain compression, P_{1dB}. Express your result in dBm.

   (b) Suppose that two signals are input to the 2-port, i.e., that $v_i = a_1 \cos\omega_1 t + a_2 \cos\omega_2 t$. List the frequencies of all components that will appear at the output of the 2-port.

   (c) Now assume that the two signals in part 3b have equal amplitudes, i.e., $a_1 = a_2 = a$. Find the maximum input signal power (for each signal) that will cause the in-band third-order components to be down 30 dB with respect to the desired signals at the output of the 2-port. You may assume that gain compression can be ignored at this input power level. Express your result in dBm.

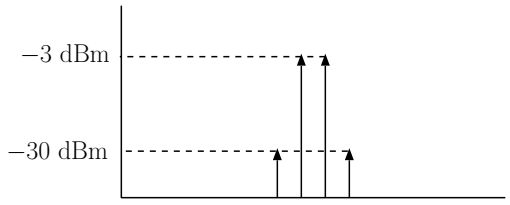


Figure 12.12: Spectrum analyzer display.

4. Two signals at closely spaced frequencies f_1 and f_2 are applied to the input of an amplifier. The input power for each of the signals is -20 dBm. The display in Figure 12.12 is seen when the output of the amplifier is connected to a spectrum analyzer:

   (a) Find the input intercept $P_I^{(i)}$ for this amplifier. Assume that gain compression is not important.

   (b) What is the operating power gain of the amplifier (in dB)?

5. Consider the receiving system in Figure 12.13:

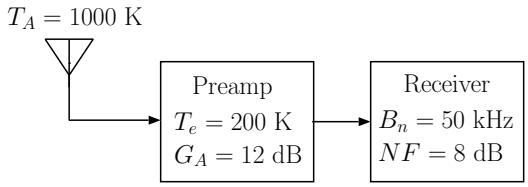


Figure 12.13: Receiving system.

   (a) Assume that the antenna and 2-ports are matched to 50 Ω. Find the available signal power level from the antenna (in dBm) that is required to give a 15 dB SNR at the output of the receiver.

   (b) The two-tone third-order input intercept level for the receiver is +6.5 dBm. Assume that the preamplifier is linear. Find the spurious-free dynamic range (DR) of the system if a 15 dB SNR is required at the output for detection. Express your result in dB.

6. A receiver has effective input temperature $T_e = 1500K$ and noise bandwidth $B_n = 3\,\text{kHz}$. The antenna temperature is 400K. The two-tone third-order input intercept for the receiver is -5 dBm. A signal-to-noise ratio of 6 dB is required for detection.

   (a) Compute the MDS for the receiver. Give your result in dBm.

   (b) Compute the spurious-free dynamic range of the receiver. Give your result in dB.

   (c) Suppose that the MDS computed in part 6a is too high for the intended application, i.e., suppose that the required MDS is -130 dBm. To achieve this

specification, we decide to add a preamplifier at the input of the receiver. Suppose that the preamplifier will have an effective input temperature $T_e = 175K$. How much gain should the preamplifier have? Specify the value that will result in an MDS = -130 dBm.

(d) Use your result from part 6c and compute the spurious-free dynamic range of the preamp-receiver combination. You may assume that the preamplifier is a linear 2-port.

7. Consider a receiver with effective input temperature T_e, two-tone third-order input intercept level $P_I^{(i)}$, and equivalent noise bandwidth B_n. Suppose we put a passive, matched attenuator in front of the receiver. The attenuator has loss $L > 1$. The physical temperature of the attenuator is equal to standard temperature, i.e., $T_{att} = T_o$. The addition of the attenuator will cause the cascade to have a higher input intercept level than that of the original receiver. The new input intercept level for the cascade will be $P_I^{(i)}L$. Denote the source temperature by T_s. Depending upon the value of T_s, the addition of the attenuator will either increase or decrease the dynamic range of the system relative to that of the original receiver. What constraint must be satisfied by T_s in order for the addition of the attenuator to increase the dynamic range of the system?

8. An amplifier is found to have the following nonlinear input-output voltage characteristic:

$$v_o = k_1 v_i + k_2 v_i^2 \tag{12.58}$$

where $k_1 = 15$ and $k_2 = 1$. This characteristic does not include a third-order term, $k_3 = 0$, so the two-tone third-order input intercept power is infinite. Define a two-tone second-order input intercept power, $P_{I2}^{(i)}$, which is the input power for each of the equal amplitude tones that causes the output power in the second-order intermodulation products at frequencies $f_1 + f_2$, $|f_1 - f_2|$ to be equal to the output power in the desired tones. Assume that the input impedance of the amplifier is 50 Ω and find $P_{I2}^{(i)}$ (in dBm).

9. Two signals at closely spaced frequencies f_1 and f_2 are applied to the input of a nonlinear amplifier. Denote the input power of the tones by P_1 and P_2, respectively. Suppose that P_1 is decreased by 2 dB. Specify the change in the output power of each of the frequency components: f_1, $2f_2$, $2f_2 - f_1$, $2f_1 - f_2$, $3f_1$. You may assume that P_1 and P_2 are small enough so that gain compression can be ignored.

10. Suppose that an AM broadcase receiver uses a nonlinear preamplifier described by:

$$v_o = k_1 v_i + k_2 v_i^2 + k_3 v_i^3 \tag{12.59}$$

and a 4-diode doubly-balanced switching-type mixer to receive a strong signal with carrier frequency 1200 kHz. The 4-diode doubly balanced mixer produces terms with frequencies $|f_{RF} \pm n f_{LO}|$, n=1,3,5... at its output. The radio uses a 455 kHz IF, the local oscillator tunes from 995-2155 kHz, and the receiver's tuning dial spans 540 - 1700 kHz. Find and list all settings of the receiver tuning dial at which you could potentially detect a signal from the station at 1200 kHz.

Chapter 13

Phase-locked Loops (PLLs)

13.1 PLL Fundamentals

Figure13.1 shows a linear model for a PLL. In this idealized linear model, the input is the phase of the reference signal and the output is the phase of the VCO. Both the reference and VCO signals are assumed to be cosinusoidal signals with constance amplitude (for now) and slowly changing phase. When the loop is locked and in a steady state, the VCO (output)

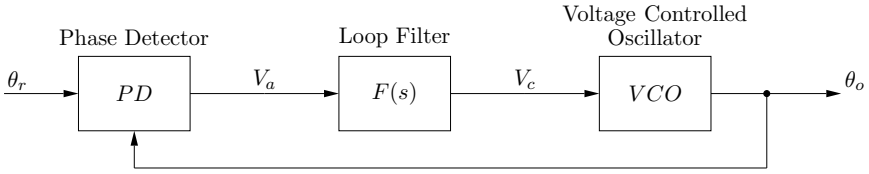


Figure 13.1: Linear model for a PLL.

frequency and reference frequency are equal, i.e.,

$$V_r(t) = V_r \cos(\omega_r t + \theta_r) \tag{13.1}$$
$$V_o(t) = V_o \cos(\omega_r t + \theta_o). \tag{13.2}$$

The phase detector (PD) produces an output voltage that is proportional to the phase error, θ_e, which is defined to be the difference between the phase of the reference signal and the VCO signal:

$$\begin{aligned} V_a(t) &= K_d\,(\theta_r(t) - \theta_o(t)) \\ &= K_d\,\theta_e(t) \end{aligned} \tag{13.3}$$

or in the s-domain:

$$\begin{aligned} V_a(s) &= K_d\,(\theta_r(s) - \theta_o(s)) \\ &= K_d\,\theta_e(s) \\ K_d &= \text{ PD gain constant (V/radian)} \end{aligned} \tag{13.4}$$

The PD output voltage is applied to the loop filter to produce the VCO control voltage, V_c

$$V_c(s) = V_a(s)F(s) \tag{13.5}$$

where F(s) has a low-pass frequency response function.

The instantaneous frequency deviation of the VCO output signal is proportional to the control voltage:

$$\frac{d\theta_o}{dt} = K_o\, V_c(t) \tag{13.6}$$

or

$$\begin{aligned} s\,\theta_o(s) &= K_o\, V_c(s) \qquad (13.7)\\ K_o &= \text{VCO gain constant, } \frac{\text{radians}}{\text{V sec}} \end{aligned}$$

13.1.1 PLL Transfer functions

$$\frac{\theta_o(s)}{\theta_r(s)} = \frac{K_o\, K_d\, F(s)}{s\, +\, K_o\, K_d\, F(s)} \tag{13.8}$$

$$\frac{\theta_e(s)}{\theta_r(s)} = \frac{s}{s\, +\, K_o\, K_d\, F(s)} \tag{13.9}$$

Sometimes the control voltage, V_c, is the desired output signal, e.g., when the loop is used as a demodulator for FM:

$$\frac{V_c(s)}{\theta_r(s)} = \frac{s\, K_d\, F(s)}{s\, +\, K_o\, K_d\, F(s)} \tag{13.10}$$

13.1.2 Loop Gain and Notation

The closed-loop transfer function $H(s)$ as defined in Equation 13.8 is

$$\begin{aligned} H(s) &= \frac{\theta_o(s)}{\theta_r(s)} \qquad (13.11)\\ &= \frac{K_o\, K_d\, F(s)}{s\, +\, K_o\, K_d\, F(s)} \end{aligned}$$

Note:

$$\begin{aligned} \frac{\theta_e(s)}{\theta_r(s)} &= \frac{\theta_r(s) - \theta_o(s)}{\theta_r(s)} \qquad (13.12)\\ &= 1 - H(s)\\ \frac{V_c(s)}{\theta_r(s)} &= \frac{s}{K_o} H(s) \end{aligned}$$

$H(s)$ can be written

$$H(s) = \frac{K_o\, K_d\, F(s)/s}{1\, +\, K_o\, K_d\, F(s)/s} \tag{13.13}$$

The quantity $A(s) = K_o\,K_d\,F(s)/s$ is called the *open loop gain*. Note that the closed loop transfer function, $H(s)$, can be written in terms of the open loop gain:

$$H(s) = \frac{A(s)}{1 + A(s)}. \tag{13.14}$$

13.1.3 Order and Type

The "order" of a PLL is defined by the highest power of s in the denominator of the *closed loop* transfer function, i.e.,

- First order:
$$H(s) = \frac{K}{s\,+\,a} \tag{13.15}$$
- Second order:
$$H(s) = \frac{K}{s^2\,+\,a\,s\,+\,b} \tag{13.16}$$

The "type" of a PLL is defined by the number of poles at the origin for the open loop transfer function, i.e.,

- Type 1:
$$A(s) = \frac{K}{s} \tag{13.17}$$
- Type 2:
$$A(s) = \frac{K}{s^2} \tag{13.18}$$

All PLL's are at least type 1, because the VCO output phase is proportional to the integral of the control voltage.

13.1.4 Loop Filters

Commonly employed loop filters fall into four classes ranging from the simplest (no filter at all) to active filters.

1. No filter:
$$F(s) = 1 \tag{13.19}$$

   Then
$$\begin{aligned} H(s) &= \frac{K_o\,K_d}{s\,+\,K_o\,K_d} \qquad (13.20)\\ &= \frac{K_o\,K_d/s}{1\,+\,K_o\,K_d/s}\\ &= \frac{\omega_L/s}{1\,+\,\omega_L/s}\\ \omega_L &= K_o\,K_d\ = \text{loop bandwidth} \end{aligned}$$

   Notice that the open loop gain $A(s) = \frac{K_o\,K_d}{s}$ has one pole at the origin, so this loop is type 1. Since the power of s in the denominator of H(s) is 1, it is a first-order loop.

2. The simplest filter that gives a second-order loop is a single-pole low-pass RC filter, as shown in Figure 13.2. If the PD output is modeled as an ideal voltage source, V_a,

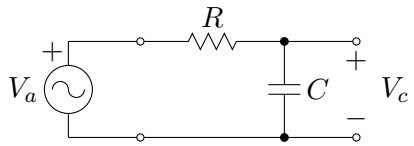


Figure 13.2: Simple low-pass loop filter.

and the control input to the VCO has high impedance, then the transfer function of the RC loop filter is:

$$F(s) = \frac{1}{1 + s\,\tau}$$

With this loop filter, the closed-loop transfer function for the PLL is

$$H(s) = \frac{1}{\frac{s^2\tau}{K_o K_d} + \frac{s}{K_o K_d} + 1} \tag{13.21}$$

This function can be written in the canonical form

$$H(s) = \frac{1}{\frac{1}{\omega_n^2}s^2 + \frac{2\zeta}{\omega_n}s + 1} \tag{13.22}$$

where

$$\omega_n = \sqrt{\frac{K_o K_d}{\tau}} \tag{13.23}$$

$$\begin{aligned} \zeta &= \frac{\omega_n}{2K_o K_d} \\ &= \frac{1}{2\sqrt{\tau K_o K_d}} \end{aligned} \tag{13.24}$$

The parameter ω_n is the "loop bandwidth" and ζ is called the "damping factor". Together the loop bandwidth and the damping factor determine how quickly the loop can respond to changes in the reference signal's frequency or phase.

3. The lag-lead filter (Figure 13.3) is often preferred to the simple low-pass RC filter because it leads to an improved phase margin and therefore a more stable loop transfer function. The lag-lead loop filter's transfer function is

$$F(s) = \frac{1 + s\tau_2}{1 + s\tau_1},$$

where

$$\tau_1 = (R_1 + R_2)C$$

$$\tau_2 = R_2 C$$

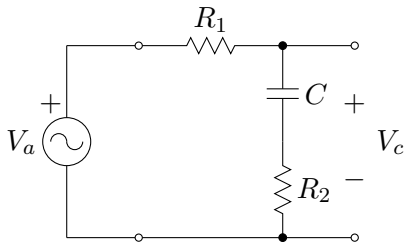


Figure 13.3: Lag-lead loop filter.

The loop transfer function is

$$H(s) = \frac{K_o K_d F(s)}{s + K_o K_d F(s)} \tag{13.25}$$

$$H(s) = \frac{K_o\, K_d (s\,\tau_2\, +\, 1)/\tau_1}{s^2\, +\, s(1\, +\, K_o\, K_d\, \tau_2)/\tau_1\, +\, K_o\, K_d/\tau_1}$$

$$H(s) = \frac{s(2\,\zeta\,\omega_n - \omega_n^2/K_o\, K_d)\, +\, \omega_n^2}{s^2\, +\, 2\,\zeta\,\omega_n s\, +\, \omega_n^2}$$

$$\omega_n = \sqrt{\frac{K_o\, K_d}{\tau_1}}$$

$$\zeta = \frac{1}{2}\left(\frac{K_o\, K_d}{\tau_1}\right)^{1/2} \left(\tau_2\, +\, \frac{1}{K_o\, K_d}\right)$$

$$= \frac{\tau_2\,\omega_n}{2}\, +\, \frac{\omega_n}{2\, K_o\, K_d}$$

4. An active filter (often used in integrated circuit PLLs) is shown in Figure 13.4: The transfer function of this filter is

$$F(s) = \frac{-A\,(s\,\tau_2\, +\, 1)}{s\,\tau_2\, +\, 1\, +\, (1\, +\, A)\, s\,\tau_1},$$

where A is the open loop gain of the op-amp. For sufficiently large A the transfer function is well approximated by

$$F(s) \simeq -\frac{s\,\tau_2\, +\, 1}{s\,\tau_1}. \tag{13.26}$$

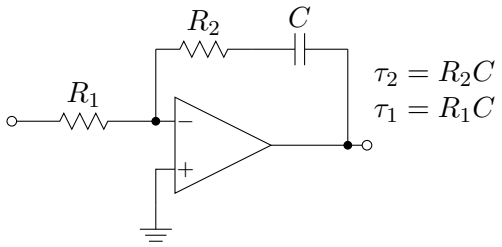


Figure 13.4: Active loop filter employing an op-amp.

Thus, for large A the loop transfer function can be written

$$\begin{aligned} H(s) &= \frac{2\,\zeta\,\omega_n\,s\,+\,\omega_n^2}{s^2\,+\,2\,\zeta\,\omega_n s\,+\,\omega_n^2} \\ \omega_n &= \sqrt{\frac{K_o\,K_d}{\tau_1}} \\ \zeta &= \frac{\tau_2}{2}\left(\frac{K_o\,K_d}{\tau_1}\right)^{1/2} \\ &= \frac{\tau_2\,\omega_n}{2} \end{aligned} \tag{13.27}$$

The high gain (large A) active loop filter and the passive lag-lead filter with large loop gain (K_oK_d) both yield open-loop transfer functions of the form

$$H(s) = \frac{2\zeta\omega_n s + \omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_{n^2}} \tag{13.28}$$

The high-gain active loop filter approximates a type 2 loop because the open loop gain has two poles at the origin:

$$\begin{aligned} A(s) &= \frac{K_oK_dF(s)}{s} \\ &\simeq -K_oK_d\frac{s\tau_2+1}{s^2\tau_1} \end{aligned} \tag{13.29}$$

The fact that the denominator of $A(s)$ contains s^2 means that there are 2 integrators in the loop. One is the VCO itself, because the phase of the VCO output signal is the integral of the control voltage ($\theta_o = \frac{K_oV_c(s)}{s}$). The other integrator is the active filter. Table 13.1 illustrates the how the parameters of the loop transfer function depend on the elements in the circuits of the passive lag-lead filter and the active filter. Figure 13.5 shows how the active filter is implemented in the Motorola MC4044 PLL integrated circuit.

13.1.5 Steady-state Error Analysis

Consider two types of inputs:

Passive lag-lead	Active Filter
$\omega_n = \sqrt{\frac{K_o K_d}{\tau_1}}$	$\omega_n = \sqrt{\frac{K_o K_d}{\tau_1}}$
$\zeta = \frac{\tau_2 \omega_n}{2}$	$\zeta = \frac{\tau_2 \omega_n}{2}$
$\tau_1 = (R_1 + R_2)C$	$\tau_1 = R_1 C$
$\tau_2 = R_2 C$	$\tau_2 = R_2 C$

Table 13.1: Comparison of passive lag-lead and active loop filters.

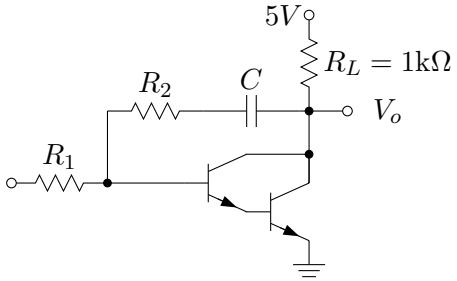


Figure 13.5: Active filter implementation in the Motorola MC4044 chip.

1. Step input in phase:

$$\theta_r(t) = \Delta\theta\, u(t) \tag{13.30}$$

2. Step input in frequency:

$$f_r(t) = f_c \,+\, \Delta f\, u(t) \tag{13.31}$$

or, equivalently:

$$\theta_r(t) = \Delta\omega\, t\, u(t) \tag{13.32}$$

The error signal $\theta_e(t) = \theta_r(t) - \theta_o(t)$ has transfer function

$$\frac{\theta_e(s)}{\theta_r(s)} = \frac{1}{1 \,+\, K_o\, K_d\, F(s)\,/s}. \tag{13.33}$$

The Laplace transform final value theorem (for stable systems) says:

$$\begin{aligned} \lim_{t\to\infty} \theta_e(t) &= \lim_{s\to 0} s\,\theta_e(s) \\ &= \lim_{s\to 0} \frac{s^2\,\theta_r(s)}{s \,+\, K_o\, K_d F(s)} \end{aligned} \tag{13.34}$$

If $\theta_r(t)$ is a step input, e.g., $\theta_r(t) = \Delta\theta\, u(t)$, then

$$\theta_r(s) = \frac{1}{s}\,\Delta\theta \tag{13.35}$$

$$\lim_{t\to\infty} \theta_e(t) = \lim_{s\to 0} \frac{s\,\Delta\theta}{s + K_o\,K_d\,F(s)}$$

Now $F(s)$ is either a constant (first-order loop) or a low-pass filter that may include poles at the origin, i.e.,

$$\lim_{s\to 0} F(s) = \frac{K}{s^n} \neq 0 \tag{13.36}$$

Thus

$$\lim_{t\to\infty} \theta_e(t) = \lim_{s\to 0} \frac{s^{n+1}\,\Delta\theta}{K_o\,K_d\,K} = 0 \tag{13.37}$$

i.e., a stable PLL will track step changes in phase with zero steady-state error. If the frequency changes suddenly, i.e.,

$$f_r(t) = f_c + \Delta f\, u(t) \tag{13.38}$$

or

$$\theta_r(t) = t\,\Delta\omega\, u(t) \tag{13.39}$$

then

$$\theta_r(s) = \frac{1}{s^2}\,\Delta\omega \tag{13.40}$$

The steady-state phase error is

$$\lim_{t\to\infty} \theta_e(t) = \lim_{s\to 0} \frac{\Delta\omega}{s + K_o\,K_d\,F(s)} \tag{13.41}$$

$$= \frac{\Delta\omega}{K_o\,K_d\,F(0)}$$

If $F(0)$ = constant (dc gain of filter = constant), then the steady-state phase error is inversely proportional to $K_o\,K_d$. Since a larger value for K_oK_d leads to a larger loop bandwidth for all the filters considered, we can conclude that a large loop bandwidth is desirable if the steady-state phase error is to be minimized.

The frequency error $f_e(t) = \frac{d}{dt}\theta_e(t)$ will tend toward zero as $t \to \infty$ (because $\theta_e \to$ constant). If it is necessary to have zero steady-state phase error in response to a step frequency input, then

$$\lim_{s\to 0} \frac{\Delta\omega}{K_oK_dF(s)} = 0 \tag{13.42}$$

or

$$\lim_{s\to 0} F(s) = \infty \tag{13.43}$$

This means that F(s) must have a pole at the origin which means that the DC gain of the filter must approach infinity. In this case the system will be type 2, since

$$A(s) = \frac{K_o\,K_d\,F(s)}{s} \tag{13.44}$$

has two poles at the origin. We can approximate the type 2 loop with a high gain active filter which can be used to achieve an essentially zero steady-state phase error. The addition of a pole at the origin can, however, cause stability problems. Stability issues will be discussed in the following section.

13.2 Stability Analysis

The stability of a loop is determined by $A(s)$ which depends on $F(s)$. A loop's stability can be studied by examining the complex-plane location of the poles of $H(s)$. This requires knowledge of the analytical form of $H(s)$. Alternatively, stability can be studied without direct knowledge of the pole locations or even an analytical description of the transfer function by using Bode plots. A Bode plot (shown in Figures 13.6 and 13.7) consists of a pair of graphs of the magnitude (in dB) and phase of the open loop gain $A(s)$ plotted on a logarithmic frequency scale. The Bode criterion for stability is simple and is equivalent to the reverse of the Barkhausen criterion for oscillation. The Bode criterion states that a loop is stable if the magnitude of $A(s)$ falls below 1 (0 dB) before the phase-shift of $A(s)$ reaches 180°.

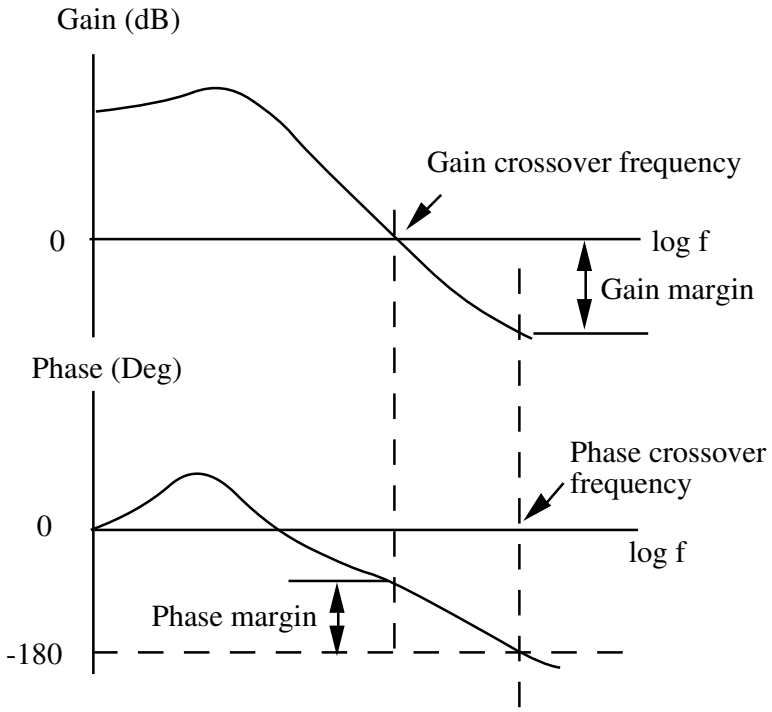


Figure 13.6: Bode plot of A(s) for a stable PLL

Summary of stability-related terminology:

- Phase margin is the difference between the actual phase and 180° at the frequency where the magnitude of the open loop transfer function is unity.

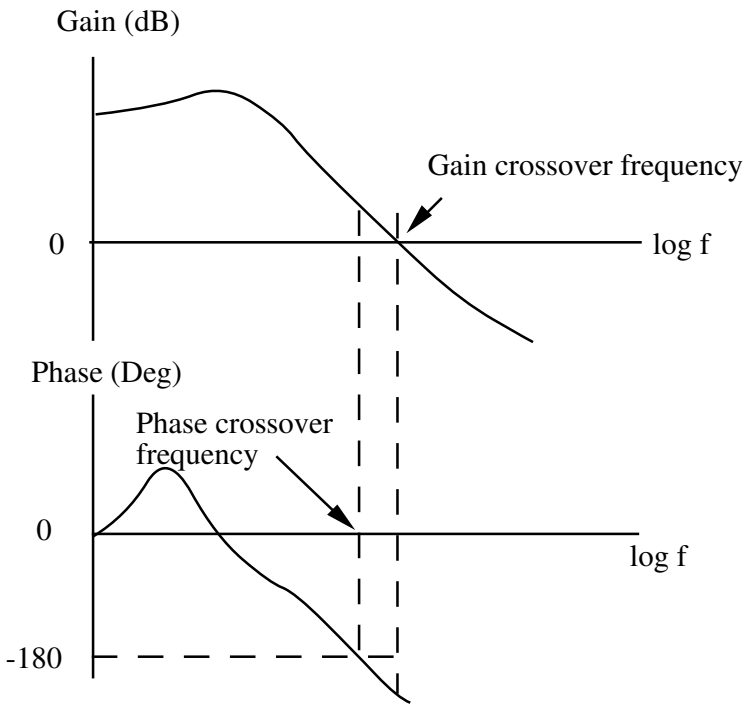


Figure 13.7: Bode plot of A(s) for an unstable PLL

- Gain margin is the number of dB below 0 dB for the gain at the frequency where the phase of the open loop transfer function is 180°.
- The greater the phase margin, the more stable the system and the more phase lag from parasitic effects can be tolerated. Phase compensation provided by the lag-lead filter can often be used to stabilize a marginally stable loop. See the examples in the following section.

13.2.1 Examples of Stability Analysis

Consider a PLL which has $K_o K_d = 50$ rad/s and which contains a simple RC low-pass filter with corner frequency $\omega_c = 100$ rad/s. The open-loop gain is

$$A(s) = \frac{K_o K_d}{s} F(s) \tag{13.45}$$

or

$$\begin{aligned} A(s) &= \frac{50}{s(1+\frac{s}{100})} \\ A(j\omega) &= \frac{50}{j\omega(1+\frac{j\omega}{100})} \end{aligned} \tag{13.46}$$

The magnitude and phase of $A(j\omega)$ are plotted in Figure 13.8. The crossover frequency,

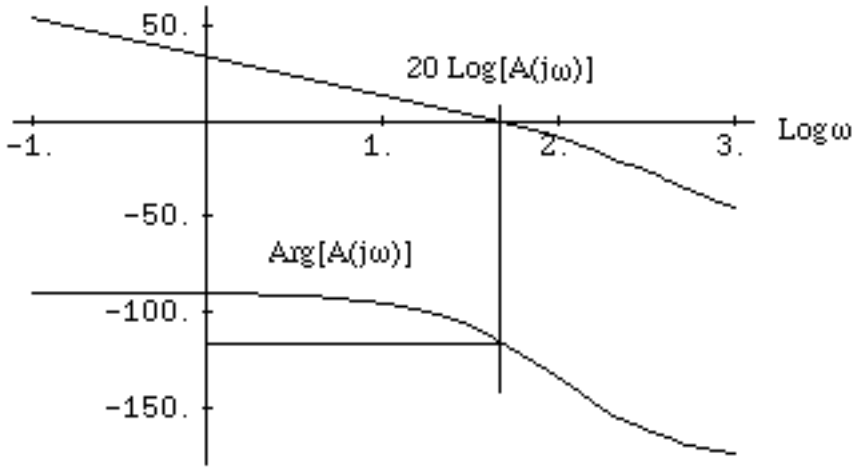


Figure 13.8: Bode plot for filter corner frequency = 100 rad/s

where $|A(j\omega)| = 1$, is $\omega = 45.51\, s^{-1}$. At this frequency the phase angle of the open-loop gain is $-114.5°$. The phase margin is therefore $180° - 114.5° = 65.6°$.

Suppose that the filter corner frequency was 10 rad/s instead of 100 rad/s. The magnitude and phase of $A(j\omega)$ are shown in Figure 13.9. In this case the gain crossover frequency is $21.3\, s^{-1}$ and the phase angle at the gain crossover is $-154.8°$. So the phase margin is

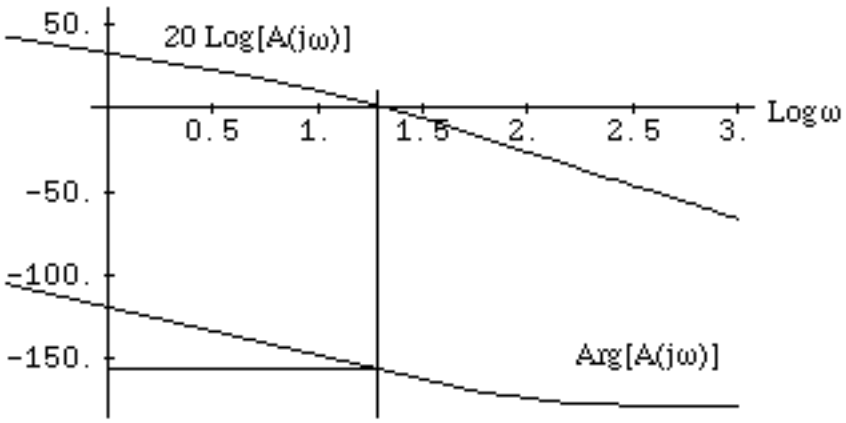


Figure 13.9: Bode plot for filter corner frequency = 10 rad/s

.

180-154.8 = 25.2°. The phase margin of 25.2° is rather small, and a relatively small perturbation of the phase angle (due to aging, temperature, or other effects) could cause the loop to become unstable and oscillate. It is therefore desirable to find a way to increase the phase margin. The phase margin can be improved by employing a lag-lead filter which adds "compensation" by adding a zero to the transfer function at 50 rad/s. The open-loop gain with the lag-lead filter is

$$A(s) = \frac{50}{s} \quad \frac{s/50 + 1}{s/10 + 1} \tag{13.47}$$

Now the gain crossover occurs at 22.4 s^{-1} and the phase angle at this frequency is -131.8° which corresponds to a phase margin of 48.2°. The effect of adding phase compensation is to cause the phase curve to turn back up away from -180° at high frequencies, thus increasing the phase margin.

13.3 Transient Response of PLLs

Consider a first-order loop and suppose that the reference phase undergoes a step change at $t = 0$:

$$\begin{aligned} \theta_e(t) &= \Delta\theta u(t) \\ \theta_o(t) &= \Delta\theta(1 - e^{-K_o K_d \tau}) \end{aligned} \tag{13.48}$$

Figure 13.11 shows how the VCO output signal responds to the step change in a first-order loop. When the loop bandwidth $K_o K_d$ is large, the time constant τ is small, and the loop responds quickly to changes in the reference signal's phase. Now, suppose that the reference signal's frequency undergoes a sudden step change. In that case:

$$\begin{aligned} \theta_r(t) &= \Delta\omega u(t) \\ \theta_o(t) &= \Delta\omega t - \tfrac{\Delta\omega}{K_o K_d}(1 - e^{-K_o K_d \tau}) \end{aligned} \tag{13.49}$$

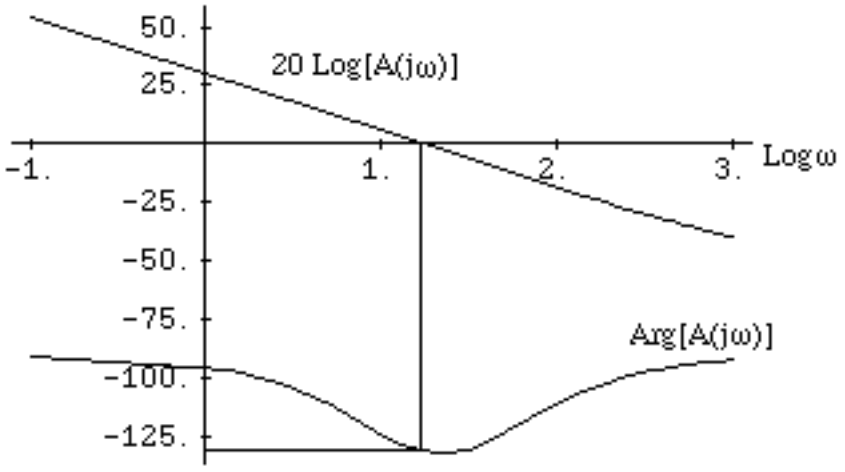


Figure 13.10: Bode plot for lag-lead filter.

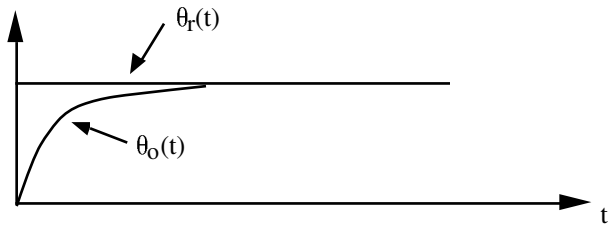


Figure 13.11: Time constant $\tau = \frac{1}{K_o K_d}$.

Figure 13.12 shows the effect the response of the VCO output phase compared to the reference phase after a frequency step. Notice that in steady-state, the VCO output phase

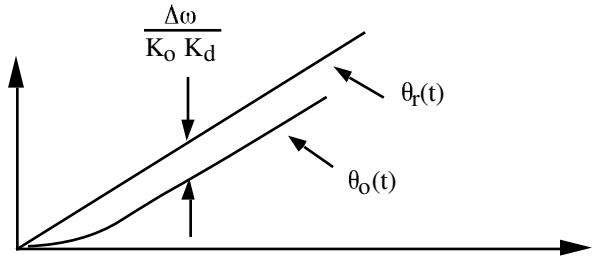


Figure 13.12: Reference signal phase and VCO output phase after a step change in the reference frequency.

lags the reference signal phase by $\frac{\Delta\omega}{K_o K_d}$. This means the the phase error is non-zero in steady-state. If the loop gain, $K_o K_d$, is large then the phase error will be small.

We will now consider the transient response of second order type-2 loops. The transfer function of such a PLL is:

$$H(s) = \frac{2\zeta\omega_n s + \omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{13.50}$$

where ω_n is the loop bandwidth (or natural frequency), and ζ is the damping factor. Both are important in determining the frequency response and transient response of the loop. Figure 13.13 shows the frequency response $|H(j\omega)|$ for different damping factors.

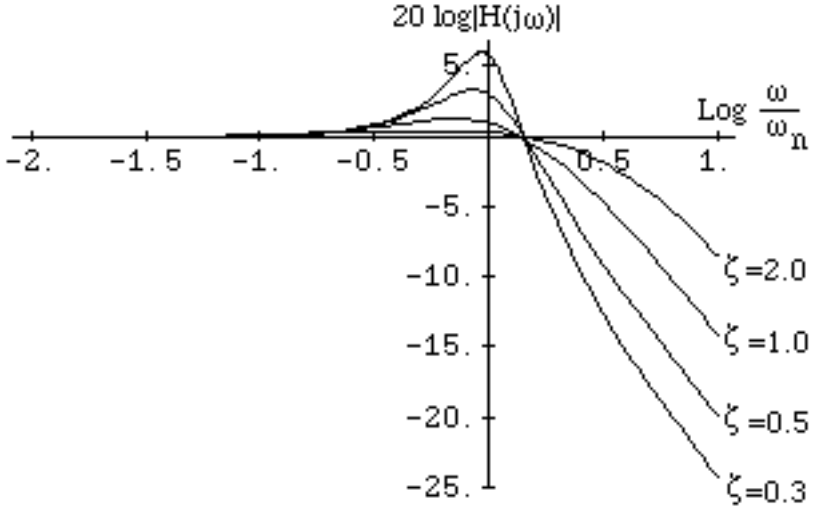


Figure 13.13: Frequency response $|H(j\omega)|$ for various damping factors

The way in which a second-order loop tracks a step change in phase or frequency can be

studied by computing the transient behavior of the phase error. Recall from Equation 13.12 that the phase error transfer function is given by:

$$\frac{\theta_e(s)}{\theta_r(s)} = 1 - H(s) \tag{13.51}$$

For a step change in the reference phase:

$$\theta_r(t) = \Delta\theta\, u(t) \tag{13.52}$$

or

$$\theta_r(s) = \frac{\Delta\theta}{s} \tag{13.53}$$

Thus

$$\theta_e(s) = \frac{\Delta\theta s}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{13.54}$$

where s has been normalized such that $s = j\omega/\omega_n$.

It is not hard to show that the transient response is of the form:

$$\theta_e(t) = \Delta\theta\, e^{-\zeta\omega_n t}\{\cosh[\sqrt{\zeta^2-1}\omega_n t] - \frac{\zeta}{\sqrt{\zeta^2-1}}\sinh[\sqrt{\zeta^2-1}\omega_n t]\} \tag{13.55}$$

The transient response is plotted in Figure 13.14 for several different values of the damping factor. Notice that (1) Increasing the damping factor results in faster settling and less ringing, and (2) the phase error tends to zero for large times. The latter was predicted earlier in our analysis of Equation 13.41.

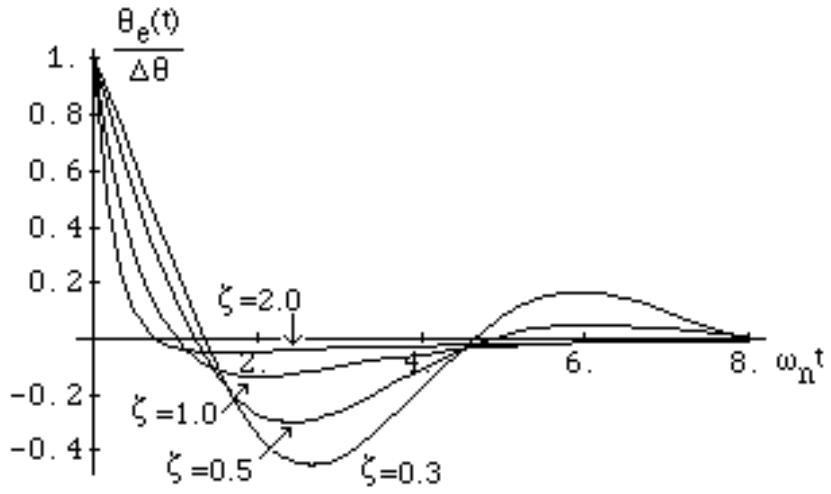


Figure 13.14: Transient responses for damping factor values

It is useful to point out that these results can also be interpreted as the transient *frequency response* due to a step change in *frequency*. To illustrate, suppose that the input is a step change in frequency, i.e.,

$$\omega_r = \omega_c + \Delta\omega\, u(t) \tag{13.56}$$

The transfer function relating VCO output frequency to the reference frequency is

$$\begin{aligned} \frac{\Delta\omega(s)}{\omega_r(s)} &= \frac{\theta_o(s)/s}{\theta_r(s)/s} \\ &= \frac{\theta_o(s)}{\theta_r(s)} \end{aligned} \tag{13.57}$$

This means that the plot given in Figure 13.14 can be interpreted as the VCO frequency response to a step change in the reference frequency.

Now consider what happens to the output *phase* when the reference frequency undergoes a step change. Suppose that initially the loop is locked at its center frequency, ω_c. At t = 0 the reference frequency is increased by an amount $\Delta\omega$, i.e.,

$$\omega_r = \omega_c + \Delta\omega\, u(t) \tag{13.58}$$

In terms of the reference phase, this input can be written as

$$\frac{d\,\theta_r(t)}{dt} = \Delta\omega\, u(t) \tag{13.59}$$

Thus

$$\theta_r(s) = \frac{\Delta\omega}{s^2} \tag{13.60}$$

The phase error is therefore

$$\theta_e(s) = \frac{\Delta\omega}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{13.61}$$

The transient response is

$$\theta_e(t) = \frac{\Delta\omega}{\omega_n}\, e^{-\zeta\omega_n t} \frac{1}{\sqrt{\zeta^2 - 1}} \sinh[\sqrt{\zeta^2 - 1}\omega_n t] \tag{13.62}$$

This function is plotted in Figure 13.15. Notice that because the loop being considered is a type 2 loop, the phase error tends to zero for large times. For type 1 loops the phase error would tend toward a constant value.

13.3.1 Summary of Second-order Loops

There are 2 parameters to play with if the loop is second-order:

$\omega_n \Rightarrow$ natural frequency

$\zeta \Rightarrow$ damping factor

Rules of thumb:

Large $\omega_n \Rightarrow$ small time constant, fast response

Large $\zeta \Rightarrow$ damped response, no ringing

Small $\zeta \Rightarrow$ ringing

Also note that:

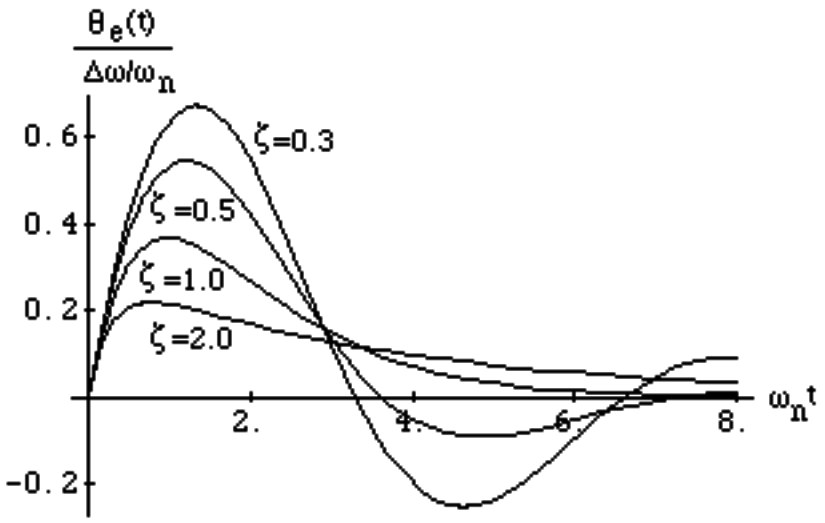


Figure 13.15: Transient response for a step change in the reference frequency.

- For a second-order loop the phase margin increases with increasing ζ.
- The natural frequency, ω_n, and damping factor, ζ, cannot be specified independently if the loop filter is a simple low-pass filter. If an active filter or lag-lead filter is employed then ω_n and ζ can be specified independently.

13.4 Applications

13.4.1 Demodulation of an FM signal

Suppose an FM signal is applied to the reference input of a PLL. The loop will try to track the deviations of the input frequency. If the loop is able to follow the deviations of the input frequency, then the control voltage V_c will be proportional to the instantaneous frequency deviation of the reference signal. In this application the PLL is called a modulation tracking loop. The transfer function relating the control voltage $V_c(s)$ to the signal phase $\theta_r(s)$ as given earlier in Equation 13.10 is

$$\frac{V_c(s)}{\theta_r(s)} = \frac{s\,K_d\,F(s)}{s\,+\,K_o\,K_d\,F(s)} \tag{13.63}$$

The input signal for FM is of the form

$$V_r(t) = V_r\,\cos\left(\omega_c t\,+\,\int_o^t m(t')\,dt'\right) \tag{13.64}$$

i.e.,

$$\theta_r(t) = \int_o^t m(t')\,dt' \tag{13.65}$$

so

$$\theta_r(s) = \frac{1}{s} M(s) \tag{13.66}$$

Thus

$$\begin{aligned} V_c(s) &= \frac{K_d F(s)}{s + K_o K_d F(s)} M(s) \\ &= \frac{1}{K_o} H(s) M(s) \end{aligned} \tag{13.67}$$

The control voltage response is a scaled and filtered version of m(t).

For frequency demodulation, the detailed shape of $|H(j\omega)|$ is important. Usually, a "flat" frequency response is desired. This means that a Butterworth-type transfer function is desirable. A second-order loop with $\zeta = 0.707$ provides the maximally flat Butterworth response, as shown in Figure 13.16.

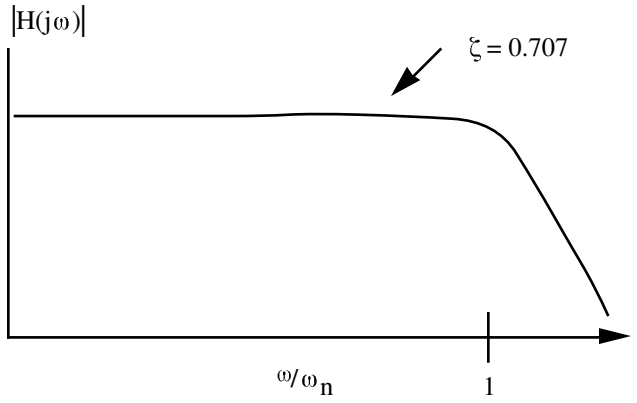


Figure 13.16: Butterworth response is obtained when $\zeta = \frac{1}{\sqrt{2}}$.

The loop bandwidth must be at least as large as the bandwidth of the modulation signal, m(t). The best results are obtained when the loop bandwidth is substantially larger than the modulation bandwidth. (See *Phaselock Techniques*, F.M. Gardner, J. Wiley, 2nd ed., 1979, Ch. 9.)

13.4.2 PLL Response to AM

Let $V_r(t) = V_r\, m(t) \cos(\omega_r t + \theta_r)$. To simplify the analysis, suppose that the PD is implemented with an ideal multiplier so that the error voltage $V_a(t)$ is

$$V_a(t) = K_d\, m(t) \sin \theta_e \tag{13.68}$$

The average value of $V_a(t)$ is

$$< V_a(t) > = < m(t) > K_d \sin \theta_e \tag{13.69}$$

As long as$< m(t) >\neq 0$, the PD output (after averaging, which is carried out by the low-pass loop filter) is proportional to sin θ_e. This means that loop will lock and will track the carrier as long as a carrier is present ($< m(t) >\neq 0$). The averaging time should be long compared to the time-scale over which $m(t)$ varies.

A PLL can be used to generate a local carrier which can be used for coherent demodulation of AM, as shown in Figure 13.17.

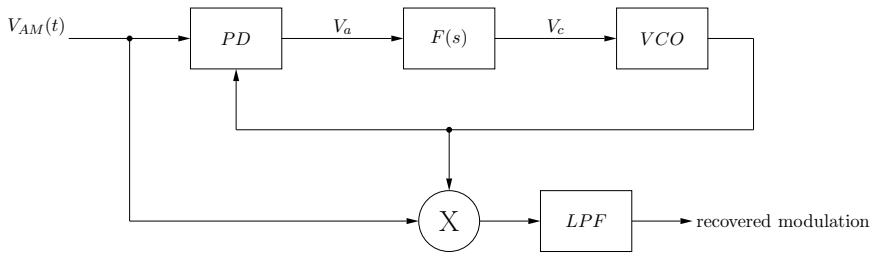


Figure 13.17: Carrier for coherent demodulation of AM

The 90° phase shift is required because the the PD is assumed to give 0 output voltage when the reference and VCO signals are in quadrature.

13.4.3 Carrier Recovery

The loop can't track the carrier if it isn't there. In DSB signals, the carrier is suppressed, so there is no carrier component for the PLL to track. There are two equivalent schemes for recovering the suppressed carrier using a PLL:

1. Squaring loop — Assume $V_r(t) = m(t)\cos(\omega_r t + \theta_r)$ and $< m(t) >= 0$ (suppressed carrier). A carrier component can be recovered if the signal is squared as in Figure 13.18: The squared input signal is $V^2_{DSB}(t) = m^2(t)\cos^2(\omega_r t + \theta_r)$. Since

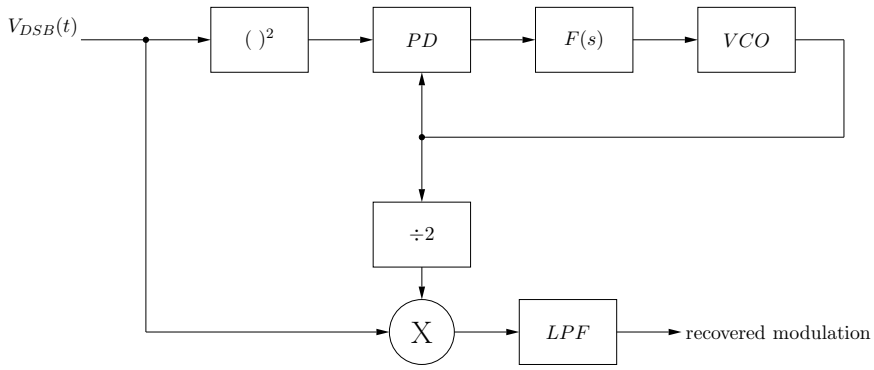


Figure 13.18: Squaring loop.

$< m^2(t) >$ is non-zero, this signal has a component at twice the carrier frequency, $2\omega_r$. The loop locks to this double-frequency component. The VCO output is divided

by 2 to produce a recovered carrier component at frequency ω_r which is then used to coherently demodulate the DSB signal. The recovered signal at the output of the divider has a 180° phase ambiguity that results from the unknown initial state of the divider. This means that the sign of the recovered modulation is arbitrary — either $+m(t)$ or $-m(t)$ could be recovered. This cause of the ambiguity can be understood by noting that the frequency divider produces a cosinusoidal signal whose argument is $\frac{1}{2}$ of the argument of the double-frequency signal presented to its input. The double frequency signal is written as $\cos(2\omega_r t + 2\theta_r + 2n\pi)$ with n any integer. The output of the divider is $\cos(\omega_r t + \theta_r + n\pi)$, which can be written as $\pm\cos(\omega_r t + \theta_r)$, depending on whether n is an even or odd integer.

2. Costas loop — refer to Figure 13.19. Without the bottom loop, this would be a

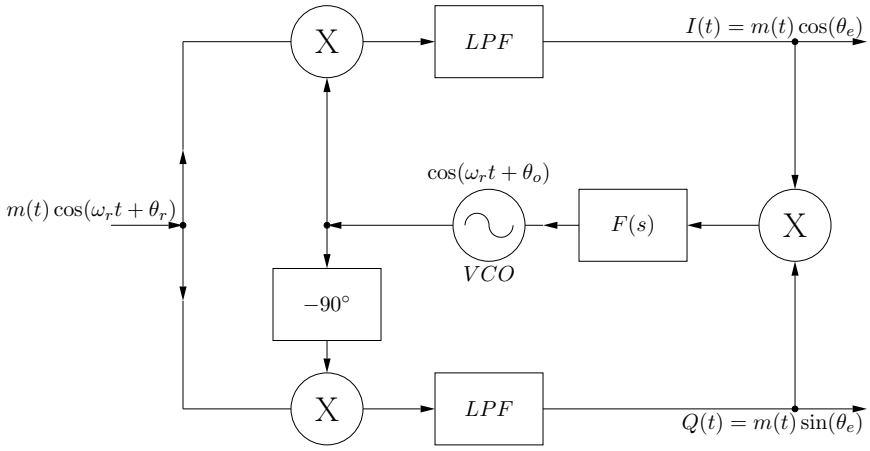


Figure 13.19: Costas loop

conventional PLL and the error signal would have an average value of zero, because $< m(t) >= 0$. With the second, or quadrature loop, the error signal is $m^2(t)\sin(\theta_r - \theta_o)\cos(\theta_r - \theta_o)$. Since $< m^2(t) >\neq 0$, the error signal is finite and operation is like the squaring loop. The Costas loop has an important advantage over the squaring loop, however, because it does not need any circuitry that operates at twice the carrier frequency. When configured as shown the Costas loop automatically provides in-phase and quadrature outputs, so the Costas loop acts as a quadrature demodulator.

13.5 Frequency Synthesis with PLL's

The simplest PLL frequency synthesizer is created by adding a frequency divider to the loop that was considered in Figure 13.1. See Figure 13.20. When the PLL is locked both of the phase detector input frequencies are equal to the reference frequency, f_r, as shown. This means that the frequency at the divider input, which is the VCO output frequency, must be Nf_r. The VCO output frequency is an integral multiple of the reference frequency. The divisor, N, of the $\div N$ divider is typically programmable. so the output frequency can be changed by changing the divisor, N. The range of available output frequencies will be

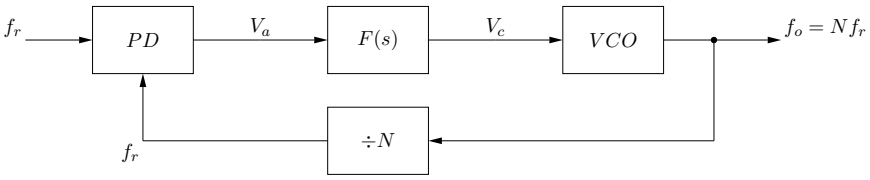


Figure 13.20: PLL frequency synthesizer.

limited by the tuning range of the VCO. The reference frequency is typically generated by a stable crystal oscillator, so when the loop is locked, the VCO output frequency will reflect the stability of the reference oscillator. If the reference oscillator frequency tolerance is δf, the VCO output frequency tolerance will be $N\delta f$.

When the divider is added to the loop, the transfer functions are modified as follows:

$$\frac{\theta_o(s)}{\theta_r(s)} = \frac{K_o\,K_d\,F(s)}{s\,+\,K_o\,K_d\,F(s)/N} \tag{13.70}$$

$$\frac{\theta_e(s)}{\theta_r(s)} = \frac{s}{s\,+\,K_o\,K_d\,F(s)/N} \tag{13.71}$$

The high gain second-order loop transfer function becomes

$$\begin{aligned} H(s) &= \frac{\theta_o(s)}{\theta_r(s)} \\ &= \frac{N(2\zeta\omega_n s+s^2)}{s^2+2\zeta\omega_n s+\omega_n^2} \\ \omega_n &= \sqrt{\frac{K_o K_d}{N\tau_1}} \\ \zeta &= \frac{\tau_2\omega_n}{2} \end{aligned} \tag{13.72}$$

The loop parameters ω_n and ζ are functions of N. As N is varied to tune the synthesizer to different frequencies the resulting variations in ω_n and ζ can result in significant changes in the loop dynamics. It is necessary to perform a worst case design analysis, i.e., using N_{min} and N_{max}, find ω_{nmin}, ω_{nmax}, ζ_{min}, ζ_{max}, and make sure that transient response and rejection of spurious signals will be adequate at all frequencies of interest.

13.5.1 Noise and Spurious Signals

In practice, both the reference signal and the VCO signal will be accompanied by noise and/or additional spurious signals. For example, to model the effect of an imperfect VCO signal which contains phase noise, the model shown in Figure 13.21. The noise on the VCO signal's phase is represented by the noise signal ϕ_{No}. The transfer function relating this

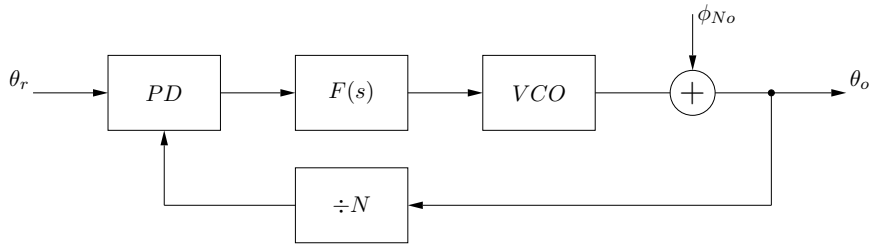


Figure 13.21: Linear model including VCO phase noise.

noise to the output phase is

$$\begin{aligned} \frac{\theta_o(s)}{\phi_{No}(s)} &= \frac{1}{1 + K_o K_d F(s)/N_s} \\ &= \frac{s}{s + K_o K_d F(s)/N_s} \end{aligned} \tag{13.73}$$

This is a high-pass frequency response function — slow, low frequency changes in the VCO output phase will be filtered out by the loop and will not reach the output, but fast, high frequency changes in the VCO output phase will be passed through to the output. This result can be understood by considering that slow changes in the VCO output phase produce a slowly varying error voltage which is passed by the loop filter and sent to the VCO control input. The error signal tunes the VCO to compensate for the original change. On the other hand a fast, high frequency, change in the VCO phase produces a high frequency error signal which does not pass through the loop filter. High frequency phase perturbations will not be filtered out by the loop because the loop cannot follow them. So the high-frequency phase perturbations appear on the VCO output signal.

The 3 major noise sources in a PLL are:

- ϕ_{Nr} — the phase noise on the reference signal
- ϕ_{Nd} — noise or spurious signals at the output of the PD. This includes harmonics of the reference signal, which are present in the output of most PD's. These harmonics will modulate the VCO and produce unwanted discrete sidebands in the VCO's frequency spectrum.
- ϕ_{No} — the intrinsic VCO phase noise, which has already been discussed.

Figure 13.22 shows a block diagram that includes these three noise sources.

The total output noise signal is

$$\phi_N = (\phi_{Nr} + \phi_{Nd}) \frac{K_o K_d F(s)}{s + K_o K_d F(s)/N} + \phi_{No} \frac{s}{s + K_o K_d F(s)/N} \tag{13.74}$$

The loop functions as a low-pass filter for phase noise on the reference signal and for reference frequency harmonics at the output of the phase detector. The loop functions as a high-pass filter for VCO phase noise. A typical VCO noise spectrum is shown in Figure 13.23. To minimize the VCO noise contribution, the loop bandwidth should be as wide as possible,

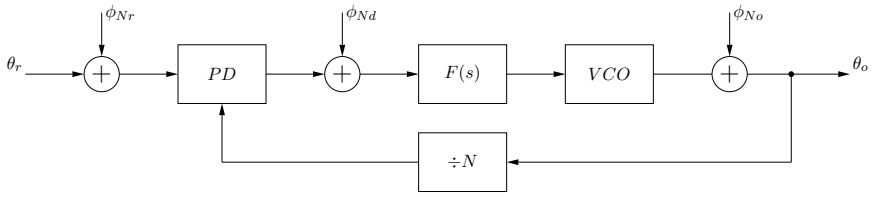


Figure 13.22: Linear model for PLL with noise sources.

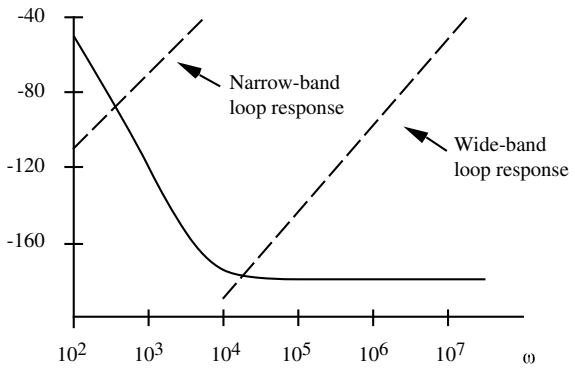


Figure 13.23: Typical VCO noise spectrum

because a large loop BW means the high-pass characteristic that filters the VCO noise will have smaller values at low frequencies. On the other hand, the loop bandwidth should be significantly less than the reference frequency in order to reduce spurious modulation of the VCO due to the reference frequency and its harmonics, which are present in the PD output. These requirements may conflict if the reference frequency is too small.

The level of the sidebands induced on the VCO output due to reference frequency components can be estimated using the following equation:

$$\frac{\text{sideband level}}{\text{carrier level}} \cong \frac{V_{ref}\, K_o}{2\,\omega_{ref}} \tag{13.75}$$

where V_{ref} is the peak voltage value of spurious frequency at the VCO input. V_{ref} is related to the spurious voltage at the output of the phase detector by

$$V_{ref} \quad = \quad V_\phi\, F(s)|_{s=j\,\omega_{ref}} \tag{13.76}$$

Suppose that the loop filter is the active filter shown in Figure 13.24 and assume that the op-amp open loop gain is very high. The loop filter transfer function will be

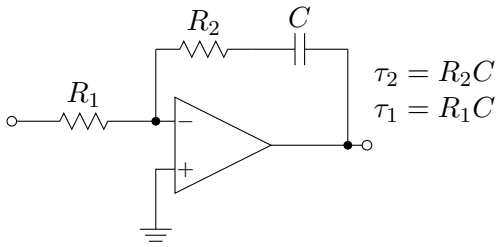


Figure 13.24: High-gain active loop filter.

$$F(s) \simeq -\frac{s\,\tau_2 + 1}{s\,\tau_1} \tag{13.77}$$

$$\tau_2 = R_2\,C \tag{13.78}$$

$$\tau_1 = R_1\,C \tag{13.79}$$

Usually, ω_{ref} is $> \frac{1}{\tau_2}$, so

$$F(j\omega_{ref}) \simeq -\frac{\tau_2}{\tau_1} = -\frac{R_2}{R_1} \tag{13.80}$$

or, in terms of loop parameters,

$$F(j\omega_{ref}) \simeq -\frac{2\,\zeta\,N\,\omega_n}{K_o\,K_d} \tag{13.81}$$

so

$$|V_{ref}| \simeq |V\phi|\frac{2\,\zeta\,N\,\omega_n}{K_o\,K_d} \tag{13.82}$$

If equation 13.82 is inserted into Equation 13.75:

$$\frac{\text{sideband level}}{\text{carrier level}} \simeq V_\phi\frac{\zeta\,N\,\omega_n}{\omega_{ref}\,K_d} \tag{13.83}$$

Usually ω_{ref}, N and K_d are predetermined by other constraints. Only ω_n and, to a lesser extent, ζ, can be adjusted to diminish the reference frequency sidebands. If the VCO is noisy, it may not be feasible to make ω_n very small because the VCO phase noise may then become objectionable. And ζ cannot be much smaller than 0.5 without running into phase margin problems and excessive ringing. In some cases it may be helpful to add additional poles to the loop filter to attenuate the reference frequency and its harmonics.

13.5.2 Phase Detectors - Digital

13.5.2.1 Exclusive-OR Phase Detector

Figure 13.25 shows an exclusive-OR phase detector. The two input signals represent the reference and VCO signals after conversion to logic-level signals. The output of the exclusive-OR gate is high if, and only if, one of the two inputs is high. In the figure the phases of the input signals differ by one quarter of a cycle, so $\theta_e = \pi/2$. The output signal consists of a

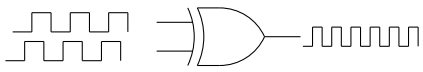


Figure 13.25: Exclusive-OR phase detector

pulse stream with 50% duty cycle and twice the frequency the input signals. If the phase difference decreases (increases) from the nominal value of $\pi/2$, the width of the output pulses will decrease (increase). When the phase difference is equal to π the input signals are always different, and the output is constantly high, and when the phase difference is zero, the input signals are always the same and the output is always zero. The lowpass loop filter averages the phase detector output waveform over many cycles, so the output of the loop filter will be approximately equal to the mean of the ex-OR output waveform. Hence, only the mean of the output waveform is important for determining the loop dynamics. The average PD output value versus phase shift θ_e is shown in Figure 13.26.

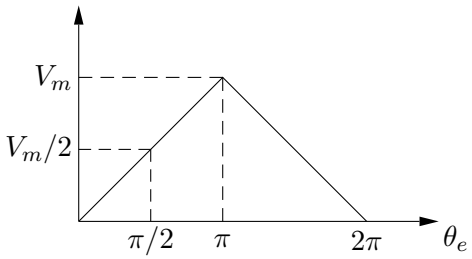


Figure 13.26: Average output versus phase shift

The exclusive-OR has an operating range of $\pm\,\pi/2$. The center of the operating range is at $\theta_r - \theta_o = \theta_e = \pi/2$, i.e., the loop locks with the reference and VCO signals in quadrature. If the phase error ever enters the region between π and 2π where the slope of the PD characteristic is negative, the loop will lose lock because the error voltage drives the VCO in the wrong direction, tending to increase the error signal rather than decrease it.

Ideally, the output contains no energy at the input frequency, but it has significant components at twice the reference frequency at also at multiples thereof. The second harmonic content is maximum at $\theta_e = \pi/2$ and has a peak-to-peak amplitude 1.27 times the peak-to-peak value of the phase detector characteristic. If these reference frequency harmonics are not sufficiently attenuated by the loop filter they can produce significant frequency modulation of the VCO and potentially objectionable discrete sidebands on the VCO output signal.

13.5.2.2 Hold-in Range of PLL, ω_H

The hold-in range is also called the lock range, tracking range, or synchronization range. It is the maximum reference frequency range over which the PD stays within its operating range. Suppose the PD operating range is $\pm\theta_m$ radians. The output frequency range is then

$$\pm K_d K_o F(0)\theta_m \tag{13.84}$$

or, defining $K_L = F(0)$ as the DC gain of loop filter,

$$\omega_H = K_d K_o K_L \theta_m \tag{13.85}$$

The loop will track reference frequencies over a range of $\pm\omega_H$.

13.5.2.3 Set-reset (SR) flip-flop

A set-reset flip-flop can be used as a phase detector (Figure 13.27). The input signals f_A and

Figure 13.27: S-R flip-flop

f_B consist of narrow pulses, as in Figure 13.28. Signal B resets the flip-flop, and signal A sets

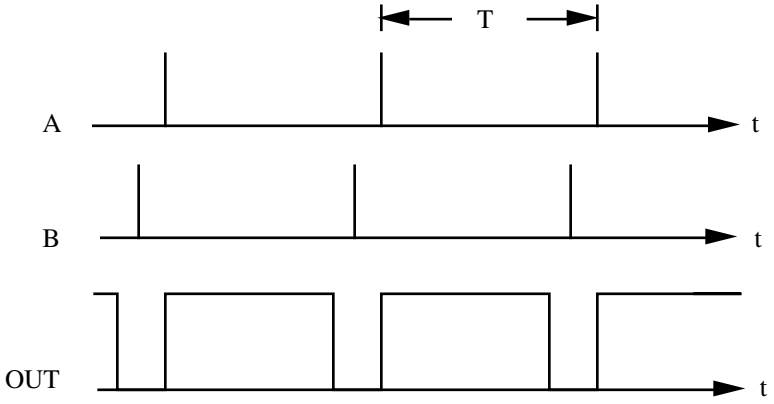


Figure 13.28: Signals f_A and f_B

it. The average output voltage as a function of phase error is shown in Figure 13.29. The SR flip-flop has twice the operating range of the exclusive-OR ($\pm\pi$). The loop locks with $\theta_r - \theta_o = \pi$. The output contains a component at the input frequency which has a maximum value at $\theta_r - \theta_o = \pi$. The amplitude of this component is 1.27 times the peak-to-peak value of the output characteristic.

Trade-offs between the exclusive-OR(EX-OR) and the S-R flip-flop are:

EX-OR	small operating range but no output at fundamental frequency, f_r
S-R	Larger operating range but output at fundamental frequency

The S-R flip-flop has some inherent problems — if the input pulses to the S-R flip-flop have finite width, there will be flat spots in the output characteristic. If the input signal goes away, the flip-flop output will remain high (or low). The loop interprets this condition

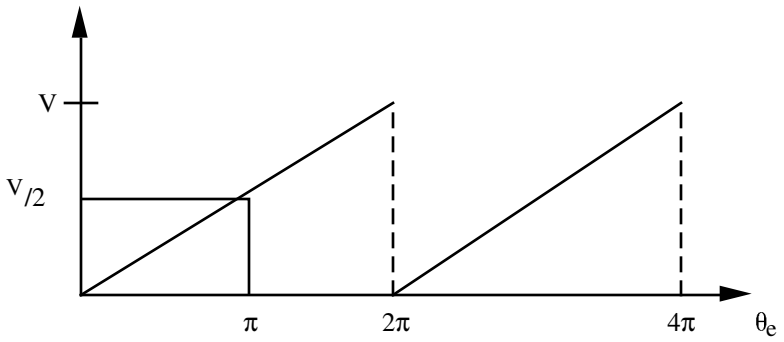


Figure 13.29: Average output as function of phase error

as a large phase error and changes the VCO frequency. Eventually the VCO will bang up against its limiting frequency and sit there.

13.5.2.4 Quad-D Phase-Frequency Detector

The quad-D phase-frequency detector (PFD) is a more sophisticated digital phase detector consisting flip-flops and additional logic. Widely available in integrated circuit form, it is called a phase-frequency detector because it provides an indication of the sign of the frequency error when the loop is out-of-lock. The PFD has two output terminals labeled U and D (for "up" and "down"). U and D can be high simultaneously, but only one can be active (low) at any time. Define the duty ratio d_u or d_D as the fractional time either terminal is active (low). The U output is only active when the phase error is greater than 0, and the U output is active when the phase error is negative. The phase characteristic when locked looks like Figure 13.30. So $d_u - d_D$ looks like Figure 13.31.

A PFD has unique properties. The active phase error range is $\pm 360°$ which is double that of the S-R flip-flop and 4 times as large as the XOR. If the loop is unlocked, only the U or the D output pulls low (active). The active output indicates the direction of the frequency error. Both outputs are quiescent at the equilibrium tracking point ($\theta_e = 0$), i.e., if the loop is locked at the center of its tracking range there are no reference frequency harmonics. In the near vicinity of equilibrium one or the other of the outputs pulls down with a small duty cycle. Thus the spurious signal is a narrow pulse at the input frequency. This is desirable, since it is much easier to filter narrow pulses than it is to filter square waves. There may be some crossover distortion in the PFD characteristic around $\theta_e = 0$. If an input signal transition is missing, the PFD interprets this as loss of lock. Since the PFD has memory, the effects propagate for more than one cycle.

The PFD has two outputs (U and D) which must be subtracted and then averaged in order to generate the VCO control voltage. Since each output can be either high or low, the difference between the two outputs has three states, positive, zero, or negative. Typically, these states are used to control a charge pump — a constant current source that can source or sink a current I_o or be in an off state. The maximum source/sink current is typically on the order of a few mA.

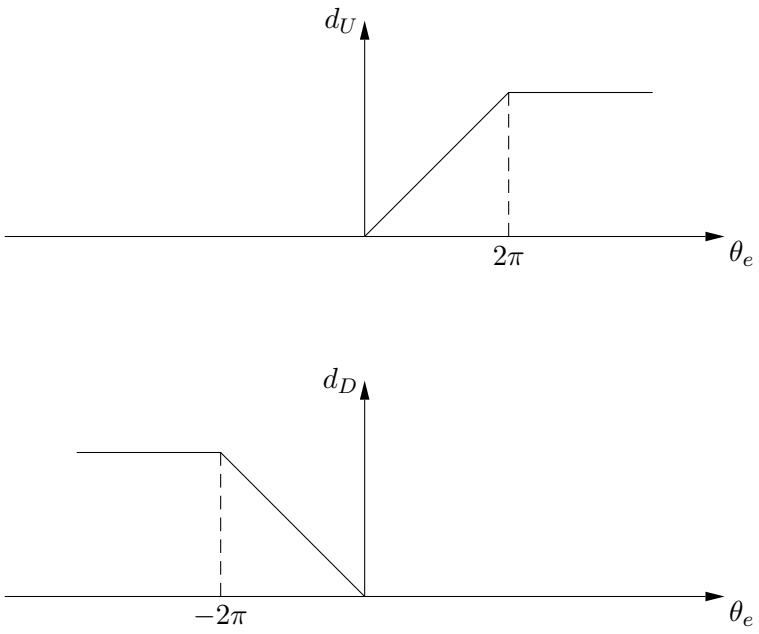


Figure 13.30: Locked phase characteristic

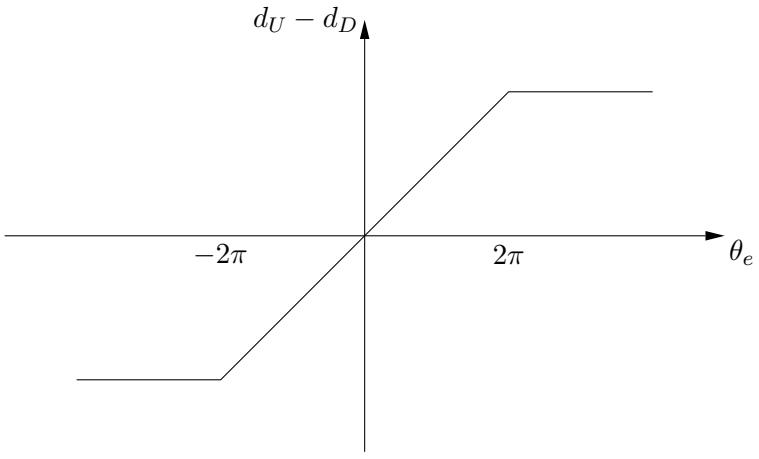


Figure 13.31: Quad-D flip-flop phase detector characteristic.

13.5.3 Examples

A PLL synthesizer has the following parameters:

$$\begin{aligned} f_{out} &= 100\,\text{MHz} \\ f_{ref} &= 100\,\text{kHz} \\ K_o &= 1\,\text{MHz/V} \Rightarrow 2\pi\,\text{x}\,10^6\,\frac{\text{rad}}{\text{V}-\text{sec}} \end{aligned} \tag{13.86}$$

The PD is an S-R flip-flop with output voltage between 0 and 5 V. Consider what the loop filter attenuation must be at 100 kHz in order to keep the reference frequency sidebands more than 40 dB down with respect to the carrier (-40 dBc) at the VCO output:

$$\begin{aligned} -40\,\text{dB} \quad &\Rightarrow \quad \frac{\text{sideband level}}{\text{carrier level}} \\ &= \quad 10^{-40/20} \\ &= \quad 0.01 \end{aligned} \tag{13.87}$$

We need 0.01 $V_\phi\,|F(j\omega_{ref})|\,\frac{K_o}{2\omega_{ref}} \leq 0.01$, where V_ϕ is the 0-to-peak spurious component voltage out of the PD. In this case

$$\begin{aligned} V_\phi &= \frac{1}{2}\,(1.27)\,(5) \\ &= \underline{3.175\ \text{V}} \end{aligned} \tag{13.88}$$

Then we need

$$\begin{aligned} |F(j\omega_{ref})| &\leq \frac{2\,\omega_{ref}}{K_o}\,\frac{1}{3.175}\,0.01 \\ &= \frac{2.10^5}{10^6}\,\frac{1}{3.175}\,0.01 \\ &= 6.30\,\text{x}\,10^{-4} \\ &= \underline{\underline{-64\,\text{dB}}} \end{aligned} \tag{13.89}$$

Suppose the loop filter is a simple RC low-pass filter with F(s) of the form

$$F(s) = \frac{1}{1 + \frac{s}{\omega_c}} \tag{13.90}$$

and the loop filter is followed by a DC amplifier, as in Figure 13.32.

So

$$\omega_c = \frac{K_o\,K_d\,K_L}{\sqrt{2}N} \tag{13.91}$$

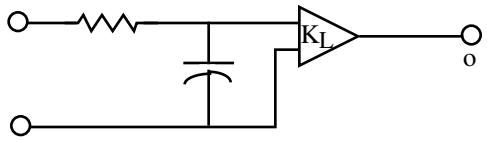


Figure 13.32: Loop filter followed by DC amplifier.

To get -40 dB sidebands, we need

$$|F(j\omega_{ref})| = 6.3\,\mathrm{x}\,10^{-4} \tag{13.92}$$

or, assuming $\omega_{ref} \gg \omega_c$

$$\frac{K_L\,\omega_c}{\omega_{ref}} = 6.3\,\mathrm{x}\,10^{-4} \tag{13.93}$$

Then, using

$$\begin{aligned} K_d &= 5\mathrm{V/cycle} \qquad (13.94)\\ &= \frac{5}{2\pi}\,\frac{V}{rad}\\ K_d &= 2\pi\,\mathrm{x}\,10^6\,\mathrm{rad/V\text{-}sec}\\ N &= 1000 \end{aligned}$$

Combining Equations 13.91 and 13.93

$$\omega_c = \sqrt{\frac{K_o\,K_d\,\omega_{ref}\,6.3\,x\,10^{-4}}{\sqrt{2}N}} = 1.183 \times 10^3 \tag{13.95}$$

or

$$f_c = 188\,\mathrm{Hz}$$

If the phase margin is to be 45°, what are the required values of ω_c, K_L, and ω_n? The open loop gain is

$$\begin{aligned} A(s) &= \frac{K_o\,K_d}{s}\,F(s)\frac{1}{N} \qquad (13.96)\\ F(s) &= \frac{KL}{1\,+\,s/\omega_c} \end{aligned}$$

so

$$\begin{aligned} A(s) &= \frac{K_o\,K_d\,K_L/N}{s(1\,+\,s/\omega_c)} \qquad (13.97)\\ A(j\omega) &= \frac{K_o\,K_d\,K_L/N}{j\omega(1\,+\,j\omega/\omega_c)} \end{aligned}$$

For a phase margin of 45°, we want

$$|A(j\omega)| = 1 \text{ when } \omega \simeq \omega_c \tag{13.98}$$

Thus

$$1 = \frac{K_o\, K_d\, K_L/N}{\omega_c \sqrt{2}} \tag{13.99}$$

From Equation 13.93:

$$K_L = \frac{\omega_{ref}}{\omega_c} \bullet 6.3 \times 10^{-4} = 0.34 \tag{13.100}$$

The natural frequency of the loop (loop bandwidth) is

$$\omega_n = \sqrt{\frac{K_o\, K_d\, K_L\, \omega_c}{N}} = 1418 \tag{13.101}$$

or

$$f_n = 226\,\text{Hz}$$

The damping factor for this loop would be

$$\zeta = \frac{1}{2}\frac{\omega_c}{\omega_\eta} \tag{13.102}$$

$$= .42$$

The settling time for the loop will be

$$\sim \frac{7}{\omega_n} = 5\,\text{ms} \tag{13.103}$$

13.5.4 Pre-scalers

A number of factors may combine to make the simple synthesizer discussed in section 13.5.3 unsuitable. For example: (i) the required channel spacing (f_r) may be too small to allow for adequate suppression of spurious components that appear at the PD output; (ii) a small f_r may result in a lock-up time which is too long; (iii) programmable dividers may not be available for operation at very high frequencies. Problem (iii) can be addressed by first dividing by a fixed-modulus divider (pre-scaler) as shown in Figure 13.33. It is possible

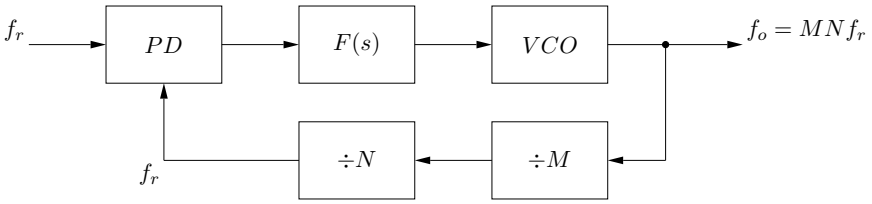


Figure 13.33: PLL frequency synthesizer with fixed pre-scaler (M) and programmable divider (N).

to obtain fixed-modulus dividers (pre-scalers) that operate well into the microwave region. The output of the $\div M$ pre-scaler will be at a lower frequency, where the fully programmable $\div N$ can operate. When the loop is locked

$$f_r = \frac{f_o}{MN} \tag{13.104}$$

or

$$f_o = N\,(Mf_r) \tag{13.105}$$

A pre-scaler allows a loop to operate at higher frequencies but the output frequency can only be changed in increments of Mf_r. To get better resolution we must decrease f_r. This will tend to make problems (i) and (ii) worse.

13.5.5 Dual Modulus Dividers

A method for obtaining good frequency resolution while operating at high output frequencies uses a variable modulus pre-scaler. Typically, the variable modulus divider is a high-speed divider with two choices for the modulus. A synthesizer using the dual modulus scheme is shown in Figure 13.34 where the high-speed divider can divide by either P or $P+Q$. The divide ratio is selected by a signal from the low-speed divider A. Here is a summary of how

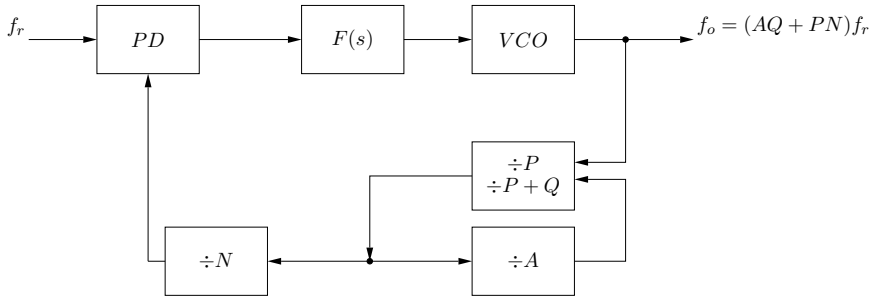


Figure 13.34: PLL synthesizer with dual modulus pre-scaler. The frequency at the inputs of the $\div N$ and $\div A$ counters is much smaller than the VCO output frequency, so these counters do not have to be very fast. The dual modulus pre-scaler does have to be fast, since it accepts the VCO output signal and must divide it by either P or $P+Q$. The dual modulus pre-scaler accepts a binary input from counter A, which instructs it to divide by P or $P+Q$. Note that N will always be greater than A.

the dual modulus pre-scaler operates:

1. Load $\div N$ and $\div A$ counters with N and A, respectively. The dual modulus pre-scaler is set to divide by $P+Q$.
   (a) After $P+Q$ pulses, both N and A counters are decremented.
   (b) Continue until the A counter reaches zero. This will occur after $(P+Q)A$ cycles of the VCO output signal, f_o.
   (c) The A counter then instructs the dual modulus pre-scaler to divide by P. The N counter contains $N-A$. (We must have $N > A$ for this scheme to work).

(d) Continue until the N counter reaches zero. This takes $P(N-A)$ cycles of f_o.

(e) The N counter outputs 1 pulse. The cycle is complete.

(f) Begin again at step (1).

The total number of f_o pulses in one complete divide cycle is

$$(P+Q)A + P(N-A) = AQ + PN \tag{13.106}$$

i.e., for every $AQ+PN$ pulses in, we get 1 pulse out to the PD. Thus, the three counters together effectively divide by $AQ+PN$. When the loop is locked

$$f_o = [AQ+PN]f_r \tag{13.107}$$

Recall that N must be greater than A for the method to work.

A frequently used divide ratio is $P=10,\ P+Q=11$. Then

$$f_o = [10N + A]\, f_r \tag{13.108}$$

Suppose that A can be set to any value from 0 to 9. Then N must be at least 10 to keep $N > A$. This means that the lowest output frequency is $100f_r$, but any integer multiple larger than $100f_r$ can be achieved. Thus the 10/11 dual modulus pre-scaler can be used to obtain integer multiples of f_r even when the output frequency is too large for a fully programmable divider to handle.

Consider an example. Suppose that we need a frequency synthesizer to cover the frequency range 100 MHz to 1099 MHz in 1 MHz increments. Since the step size is 1 MHz, it is necessary to choose $f_r \leq 1$ MHz. Let $f_r = 1$ MHz. Suppose that 1099 MHz is too fast for a fully programmable divider, however a fast ($P=10,\ P+Q=11$) dual modulus pre-scaler is available. Then we use the following parameters:

$$\begin{aligned} f_r &= 1 \text{ MHz} \\ N &= 100 \\ A &\Rightarrow 0 \text{ to } 99 \\ f_o &= [1000 + A]\, f_r \end{aligned} \tag{13.109}$$

Notice that the largest frequency at the input to the $\div N$ and $\div A$ counters will be 109.9 MHz.

Now, suppose we need to cover 1000.00 to 1000.99 MHz in 10 kHz increments. Then we need $f_r = 10$ kHz. The minimum divide ratio required would be

$$\begin{aligned} \frac{1000}{0.01} &= 10^5 \\ &= \frac{f_o\,(min)}{f_r} \end{aligned} \tag{13.110}$$

Since A must run from 0 to 99, N_{min} is 100. Parameters that would work for this situation are:

$$\begin{aligned} N &= 10000 \\ P &= 10 \\ P+Q &= 11 \end{aligned} \tag{13.111}$$

The maximum frequency at the input to the fully programmable dividers would be 0.100099 MHz in this case.

Consider an example using parameters available from a commercially available synthesizer chip. The chip has a 5-bit A counter, so that A can range from 0 to 31. The N counter is a 13-bit counter and N can be set to any value in the range 3 to 8191. The dual-modulus pre-scaler uses $P = 8,\ \ P + Q = 9$. The output frequency will be

$$f_o = (8N + A)f_r,$$

where $0 \leq A \leq 31$, $3 \leq N \leq 8191$, and $N > A$. Suppose that it is necessary to design a synthesizer that covers $230 - 240$ MHz in 100 kHz steps and we wish to select possible values for N and A in order to tune the synthesizer to a particular output frequency. For example, to tune the synthesizer to the middle of the range (235 MHz) the effective divide ratio must be $235/0.1 = 2350 = (8N + A)$. The N counter can be set to the integer part of 2350/8, which is $N = 293$. Then, $8N = 8(293) = 2344$, so the A counter must be set to $A = 6$. To tune to 239.6 MHz, $(N, A) = (299, 4)$ would work, as would $(298, 12)$ or $(297, 24)$.

13.6 References

1. Best, Roland E., *Phase Locked Loops*, McGraw Hill, 1984.

2. Egan, William F., *Frequency Synthesis by Phaselock*, J. Wiley & Sons, New York, 1981.

3. Gardner, Floyd M., *Phaselock Techniques*, (2nd ed.), J. Wiley & Sons, New York, 1979.

4. Manassewitsch, Vadim J., *Frequency Synthesizers, Theory and Design*, J. Wiley & Sons, New York, 1987.

5. Smith, Jack, *Modern Communications Circuits*, McGraw Hill, 1986.

13.7 Homework Problems

1. Consider the first-order PLL shown in Figure 13.35. The gain constants for the VCO and phase detector are $K_o = 10^7$ radians/Volt-second, and $K_d = 10$ Volt/radian, respectively. The center frequency (free-running frequency) of the VCO is $f_c = 35.0$ MHz. Assume that the loop is locked.

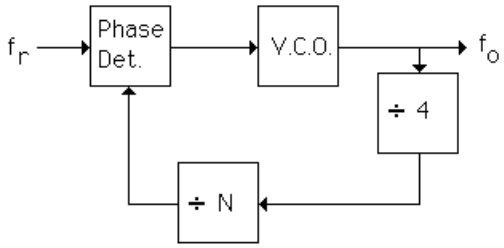


Figure 13.35: First-order PLL frequency synthesizer.

   (a) If N = 10 and the reference frequency $f_r = 1.0$ MHz, what is the output frequency, f_o?

   (b) What is the steady-state output voltage from the phase detector?

   (c) What is the steady-state phase error $\theta_e = \theta_r - \theta_o$? Give your result in degrees.

Appendix A

Circuit Models for BJT and FET

A.1 Hybrid-pi equivalent circuit for bipolar junction transistor (BJT)

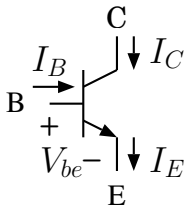


Figure A.1: BJT

The following approximate relationship is useful when the base-emitter junction is forward biased:

$$I_C = I_S \ \exp[\frac{V_{be}}{V_T}], \tag{A.1}$$

where

$$V_T = \frac{kT}{q} \simeq \ 25\text{mV at room temperature.} \tag{A.2}$$

The base current, I_b, is related to the collector current by

$$I_b = \frac{I_C}{\beta}. \tag{A.3}$$

These relationships can be used to solve for the small-signal input resistance, r_π (see Fig-

ure A.2)

$$\begin{aligned} r_\pi &= [\frac{\partial I_b}{\partial V_{be}}]^{-1}{}_{V_{be}=V_{beq}} \\ &= \frac{V_T\beta}{I_{CQ}} \\ r_\pi &\simeq \frac{.025\beta}{I_{CQ}} \end{aligned} \tag{A.4}$$

and the small-signal transconductance

$$\begin{aligned} g_m &= [\frac{\partial I_C}{\partial V_{be}}]_{V_{be}=V_{beq}} \\ &= \frac{I_{CQ}}{V_T} \\ &= \frac{\beta}{r_\pi} \\ g_m &\simeq 40 I_{CQ} \end{aligned} \tag{A.5}$$

where I_{CQ} is the quiescent collector current.

A useful linear model for the behavior of small high frequency signals superimposed on the DC bias point is the small-signal hybrid-pi model shown in Figure A.2. Typically $r_o \sim$

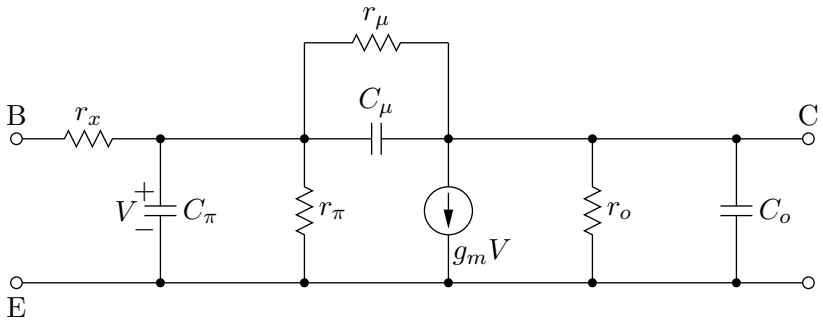


Figure A.2: Hybrid-pi small-signal model for BJT

(tens to hundreds of $k\Omega$), $r_x \sim$ (a few tens of Ω), and $r_\mu > \beta r_o$. Since r_o and r_μ are relatively large resistances, and each is shunted by a capacitance, r_o and r_μ can usually be ignored at high frequencies. On data sheets C_μ is often given as C_{ob}, which is the output capacitance in the common-base configuration. Typical values for C_μ range from a few tenths of a pF to a few pF. Data sheets may not give C_π explicitly, but will indicate the value of f_T for a particular bias current, where

$$f_T = \frac{1}{2\pi}\frac{g_m}{C_\pi + C_\mu}, \tag{A.6}$$

and f_T is the frequency where the short-circuit current gain has a magnitude of unity. The value of f_T depends on g_m and, therefore, on how the transistor is biased.

The -3dB frequency for the short-circuit current gain is denoted by f_β and is given by

$$f_\beta = \frac{1}{2\pi} \frac{1}{r_\pi (C_\pi + C_\mu)} \tag{A.7}$$

Note that the -3dB frequency and the unity gain frequency are related as follows:

$$\frac{f_T}{f_\beta} = \beta \tag{A.8}$$

For $f < f_\beta$ the short-circuit current gain has magnitude β, while for $f > f_\beta$ the short-circuit current gain is approximately f_T/f.

A.2 Hybrid-pi equivalent circuit for field effect transistor (FET)

Figure A.3 looks very similar to the BJT model (Figure A.2). The gate-source resistance, r_{gs}, is generally large compared to the impedance of C_{gs}. When this condition holds (i.e. at sufficiently high frequencies) r_{gs} can be omitted from the model.

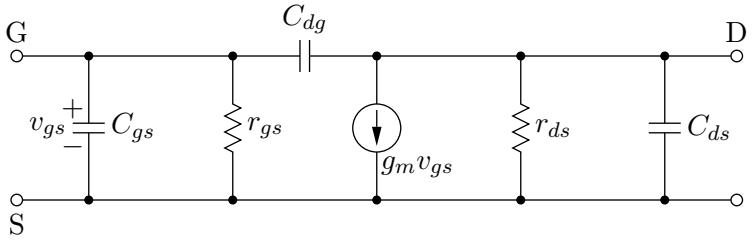


Figure A.3: Hybrid-pi equivalent circuit for the FET

For a junction FET (JFET), the transconductance is proportional to the square root of the drain current, I_D. The proportionality constant depends on the saturation drain current, I_{DSS}, and the pinchoff voltage, V_P, i.e.

$$g_m = \frac{2}{|V_P|} \sqrt{I_{DSS} I_D}. \tag{A.9}$$

The parameters I_{DSS} and V_P are usually available from device data sheets.

A.3 Large-signal transconductance of a BJT with sinusoidal V_{be}

When a BJT is driven with a sinusoidal signal such that the base-emitter voltage swing approaches or exceeds a few tens of mV, the collector current waveform becomes non-sinusoidal.

It is useful to study the effect of sinusoidal base-emitter voltage swing on the collector current waveform. The following approximate relationship between base-emitter voltage V_{be} and collector current I_C can be used as a starting point

$$I_C = I_S e^{V_{be}\, q/kT} \tag{A.10}$$

where kT/q is approximately 25mV for T=290K, and I_S is a constant. Decompose the base-emitter voltage and collector current into quiescent and time-varying components, i.e.,

$$I_C \;=\; I_{DC} \,+\, i_C \tag{A.11}$$

$$V_{be} \;=\; V_{DC} \,+\, v_{be} \tag{A.12}$$

where lower-case letters refer to the time-varying component of the quantity. Later, we will make the assumption that the transistor bias network acts to keep the DC component of the collector current at a nearly constant value. Suppose that the time-varying component of the base-emitter voltage is sinusoidal, i.e.,

$$V_{be} = V_{DC} \,+\, v_1 cos\omega t \tag{A.13}$$

and let $x = v_1 q/kT = v_1/25\text{mV}$ (at room temperature). Then

$$I_C = I_S \exp[V_{DC} q/kT] \exp[\, x\, cos\omega t] \tag{A.14}$$

The term $\exp[\, x \cos \omega t]$ is a non-sinusoidal periodic function of time and can be expanded in a Fourier series. The series is

$$\exp[\, x \cos \omega t] = I_o(x) \,+\, 2 \sum_{n=1}^{\infty} I_n(x) \cos(n\omega t) \tag{A.15}$$

where the coefficients $I_n(x)$ are values of the modified Bessel function of the first kind. Using this relationship, the collector current waveform can be written as

$$I_C = I_{DCo}\, [I_o(x) + 2 \sum_{n=1} I_n(x) \cos(n\omega t)] \tag{A.16}$$

Here I_{DCo} is the DC component of collector current when the time-varying component of the input signal is equal to zero ($v_1 = 0$). The DC component of the collector current when the time-varying component of the input signals is not zero is given by

$$I_{DC} = I_{DCo} I_o(x). \tag{A.17}$$

This function is plotted in Figure A.4, which shows that $I_o(x)$ grows very rapidly when the base-emitter voltage swing exceeds a few tens of mV. This analysis has assumed that the DC component of the base-emitter voltage is held constant. Practical amplifier and oscillator circuits will employ either constant current source bias for I_{DC} or negative feedback (by including a resistor in series with the emitter and ground) to force I_{DC} to be nearly constant. In such cases the DC component of V_{be} (denoted by V_{DC}) will decrease with increasing v_1 such that I_{DC} is more or less independent of v_1. With constant-current bias we can write:

$$I_C = I_{DC}\, [1 + 2 \sum_{n=1}^{\infty} \frac{I_n(x)}{I_o(x)} \cos(n\omega t)] \tag{A.18}$$

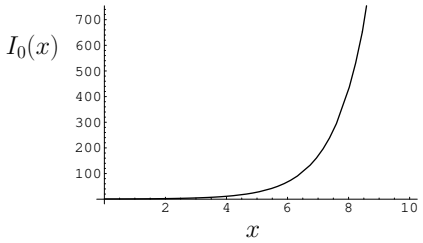


Figure A.4: The function $I_o(x)$ governs how the DC component of the collector current grows as the amplitude of sinusoidal base-emitter voltage increases ($x = v_1 q/kT$), assuming that the DC component of the base-emitter voltage is held constant.

where I_{DC} is treated as a constant. Examination of A.18 shows that the collector current waveform has components at DC, the fundamental frequency, and harmonics of the fundamental frequency. The relative amplitude of the harmonics and fundamental gives an indication of how "sinusoidal" the collector current waveform will be. Assuming constant-current bias, the amplitude of the fundamental component is proportional to $2I_1(x)/I_0(x)$. This quantity is plotted in Figure A.5. The amplitude of the fundamental approaches a constant when the base-emitter voltage swing exceeds 25 mV (i.e. for $x > 1$). The relative strengths of the second and third harmonics to the fundamental ($I_2(x)/I_1(x)$ and $I_3(x)/I_1(x)$, respectively) are also shown in Figure A.5. Notice that the harmonic amplitudes grow rapidly

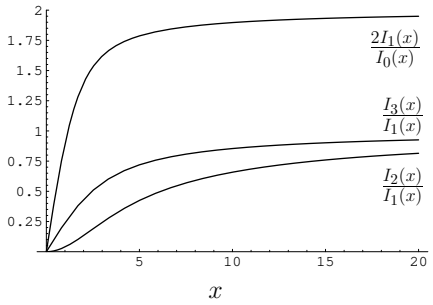


Figure A.5: $2I_1(x)/I_o(x)$ is the relative amplitude of the fundamental component of the collector current. The other curves show the ratio of second and third harmonic amplitudes to the fundamental amplitude.

when x is greater than 1 ($v_1 > 25$ mV). At large values of x the fundamental amplitude approaches twice the DC bias current, and the harmonic amplitudes approach that of the fundamental. Thus it becomes apparent that the base-emitter voltage swing must be kept as small as possible if the collector current waveform is to be sinusoidal.

This approximate analysis of the nonlinear characteristics of the BJT can also help us to gain some intuitive feeling for the saturation mechanism that limits the growth of oscillations in self-limiting oscillators, and for the reduction in apparent gain that results when a BJT amplifier is driven by a large amplitude input signal. For these purposes we consider only

the collector current component at the fundamental frequency, under the assumption that a resonant network effectively removes the harmonics. Then we can compute a *large-signal* transconductance for the transistor. With a sinusoidal input signal the transconductance is simply the ratio of the amplitudes of the fundamental component of i_C and v_1. The small-signal transconductance can be obtained in the limit as x approaches 0, i.e.,

$$g_m = \lim_{x \to 0} I_{DC} \frac{2\, I_1(x)}{v_1 I_0(x)} = \frac{I_{DC}}{kT/q} \tag{A.19}$$

At room temperature $kT/q = 25\,\text{mV}$, so

$$g_m = \frac{I_{DC}}{25\,\text{mV}} \approx 40 I_{DC} \tag{A.20}$$

The large signal transconductance is

$$G_m(x) = I_{DC} \frac{2\, I_1(x)}{v_1 I_0(x)} = I_{DC} \frac{q}{kT} \frac{2\, I_1(x)}{x\, I_0(x)} = g_m \frac{2}{x} \frac{I_1(x)}{I_0(x)} \tag{A.21}$$

The ratio of the large signal to small-signal transconductance is shown in Figure A.6. This

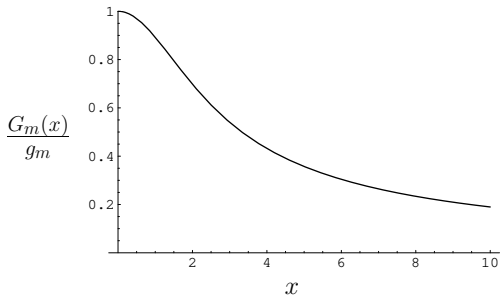


Figure A.6: Ratio of large signal to small-signal transconductance

result shows that the large signal transconductance of the transistor decreases from the small-signal value as the base-emitter voltage swing (x) increases. Thus in an oscillator application, as the oscillation amplitude grows, the effective transconductance decreases. The oscillations will continue to grow until the transconductance has been reduced to a value that causes the (large-signal) loop gain to be equal to 1. The large-signal transconductance is also important in determining the behavior of stable (non-oscillating) transistor circuits under large signal conditions. For example, the decrease of transconductance under large signal excitation is responsible for gain compression in transistor amplifiers when large signals are applied. These effects will discussed in some detail in Chapter 12.

Appendix B

Three-winding Transformer

The three-winding transformer finds many applications in RF circuits. Examples include balanced-to-unbalanced conversions ("baluns"), power splitters and combiners and voltage adders and subtractors. This chapter will explore some of these applications. Throughout this discussion it will be assumed that the transformer is ideal — in other words, we assume that the device is lossless, that the windings are perfectly coupled, and that the self-inductance of each winding is infinite. In practice, perfect coupling can be approximated by constructing all three windings on a common high-permeability core. Infinite self-inductance is a good approximation if the inductance of each winding is large compared to the impedance connected across the winding

The schematic representation for an ideal three-winding transformer is shown in Figure B.1. Two parallel lines running alongside the three windings remind us that a practical implementation of the transformer requires all three coils to be wound on a common core that has large relative magnetic permeability so that all three coils are linked by the same magnetic flux. By virtue of the perfect coupling between windings, the voltages across windings b and c can be written in terms of the voltage across winding a as follows:

$$V_b = \frac{N_b}{N_a} V_a, \quad V_c = \frac{N_c}{N_a} V_a \tag{B.1}$$

Since the transformer is assumed to be lossless, the total power delivered to the device must be zero, i.e.

$$\Re(V_a I_a^* + V_b I_b^* + V_c I_c^*) = 0,$$

where $\Re()$ takes the real part of its argument. Using the voltage relationships, the voltages can be written in terms of the voltage across any one of the windings. Using V_a the lossless condition becomes

$$\Re(V_a(I_a^* + \frac{N_b}{N_a} I_b^* + \frac{N_c}{N_a} I_c^*)) = 0.$$

The lossless condition must hold for any applied voltage, V_a, so the currents must satisfy

$$N_a I_a + N_b I_b + N_c I_c = 0. \tag{B.2}$$

Equations B.1 and B.2 are called the ideal transformer equations, and they fully characterize the ideal 3-winding transformer.

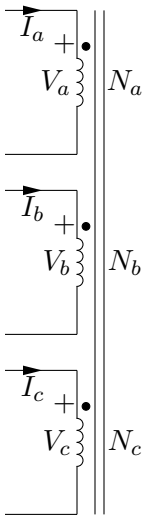


Figure B.1: Schematic representation of a three-winding transformer. The polarity of the windings is indicated by a dot. The dot convention is such that when current flows into the dot in each winding, the resulting magnetic fluxes within the core add constructively. The turns ratio is $N_a : N_b : N_c$ for windings a, b, and c, respectively.

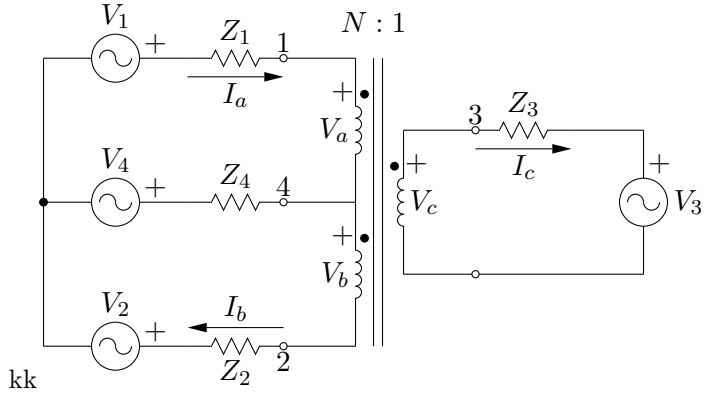


Figure B.2: The three-winding transformer configured as a 4-port device.

For simplicity's sake, we'll consider a system with $N_a : N_b : N_c = N : N : 1$, and with windings a and b connected so that they share a common terminal as shown in Figure B.2. This configuration results in a 4-port device.

Taking into account the current directions defined in Figure B.2, the ideal transformer relations can be written

$$V_a = V_b = N\,V_c \tag{B.3}$$
$$N(I_a + I_b) = I_c \tag{B.4}$$

The loop equations are

$$-V_1 + Z_1\,I_a + V_a + (I_a - I_b)Z_4 + V_4 = 0 \tag{B.5}$$

$$-V_4 + (I_b - I_a)Z_4 + V_b + Z_2\,I_b + V_2 = 0 \tag{B.6}$$

$$-V_c + I_c\,Z_3 + V_3 = 0. \tag{B.7}$$

Use B.3 andB.7 in B.5 and B.6 to eliminate V_a, V_b, V_c:

$$I_a(Z_1 + Z_4) + I_b(-Z_4) + I_c(N\,Z_3) = V_1 - V_4 - NV_3 \tag{B.8}$$

$$I_a\,(-Z_4) + I_b\,(Z_2 + Z_4) + I_c\,(N\,Z_3) = V_4 - V_2 - NV_3 \tag{B.9}$$

Equations B.4, B.8 and B.9 can be written in matrix form:

$$\begin{bmatrix} Z_1 + Z_4 & -Z_4 & N Z_3 \\ -Z_4 & Z_2 + Z_4 & N\,Z_3 \\ N & N & -1 \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \end{bmatrix} = \begin{bmatrix} V_1 - V_4 - N\,V_3 \\ V_4 - V_2 - N\,V_3 \\ 0 \end{bmatrix} \tag{B.10}$$

B.1 Conjugate Ports

We now consider some useful special cases. It is possible to select the termination impedances so that there is isolation between ports. For example, suppose it is desired to isolate port 4 from port 3 so that there will be no output at port 4 when a signal is applied at port 3. To proceed, we turn off V_1, V_2, V_4 and find the condition for no output at port 4.

The current at port 4 will be $I_a - I_b$. Using Cramer's Rule to solve for I_a:

$$I_a = \frac{D_a}{D} = \frac{\begin{vmatrix} -N\,V_3 & -Z_4 & N\,Z_3 \\ -N\,V_3 & Z_4 + Z_2 & N\,Z_3 \\ 0 & N & -1 \end{vmatrix}}{D}$$

where D is the system determinant and D_a is the determinant of the matrix formed by replacing the first column with the source column vector (the right-hand side of equation

B.10). Evaluating D_a:

$$\begin{aligned} D_a &= -N\,V_3[-(Z_4 + Z_2) - N^2\,Z_3] + Z_4[N\,V_3] + N\,Z_3[-N^2\,V_3] \qquad \text{(B.11)} \\ &= V_3\,N(2\,Z_4 + Z_2) \end{aligned}$$

Similarly, I_b is obtained from:

$$I_b = \frac{D_b}{D} = \frac{\begin{vmatrix} Z_1 + Z_4 & -N\,V_3 & N\,Z_3 \\ -Z_4 & -N\,V_3 & N\,Z_3 \\ N & 0 & -1 \end{vmatrix}}{D}$$

$$\begin{aligned} D_b &= (Z_1 + Z_4)[N\,V_3] + N\,V_3[Z_4 - N^2\,Z_3] + N\,Z_3[N^2 V_3] \qquad \text{(B.12)} \\ &= V_3\,N\,(2\,Z_4 + Z_1) \end{aligned}$$

The current flowing in the impedance Z_4 is

$$\begin{aligned} I_a - I_b &= (D_a - D_b)\,/D \qquad \text{(B.13)} \\ &= V_3\,N\,(Z_2 - Z_1)\,/D. \end{aligned}$$

The current in Z_4 will be zero if $Z_1 = Z_2$. Thus, port 4 is isolated from port 3 if $Z_1 = Z_2$. It is left as an exercise to show that the isolation works both ways, i.e. that port 3 will be isolated from port 4 when $Z_1 = Z_2$. When ports 3 and 4 are isolated, we say that ports 3 and 4 are "conjugate."

Suppose we wish to isolate port 1 from port 2. To find the necessary constraint on the terminating impedances, set $V_1 = V_3 = V_4 = 0$ and solve for I_a due to V_2. Setting $I_a = 0$ will yield the desired condition:

$$I_a = \frac{D_a}{D} = \frac{\begin{vmatrix} 0 & -Z_4 & N\,Z_3 \\ -V_2 & Z_4 + Z_2 & N\,Z_3 \\ 0 & N & -1 \end{vmatrix}}{D}$$

$$\begin{aligned} &= [Z_4\,V_2 + N\,Z_3(-N\,V_2)]\,/D \qquad \text{(B.14)} \\ &= V_2[Z_4 - N^2\,Z_3]\,/D \end{aligned}$$

Thus, port 1 will be isolated from port 2 if

$$Z_4 = N^2\,Z_3 \qquad \text{(B.15)}$$

Again, it is left as an exercise to show that condition B.15 also causes port 2 to be isolated from port 1, i.e. $I_b = 0$ with $V_1 \neq 0$ and $V_2 = V_3 = V_4 = 0$. That is, the isolation works "both ways." When ports 1 and 2 are isolated, we say that ports 1 and 2 are "conjugate."

If ports 3 and 4 are conjugate *and* ports 1 and 2 are conjugate then the system is called the biconjugate transformer.

B.2 Hybrid Transformer

Recall that the conditions for a biconjugate system are

$$Z_1 = Z_2 \tag{B.16}$$

$$Z_3 = \frac{1}{N^2} Z_4. \tag{B.17}$$

The system of equations for the biconjugate transformer is

$$\begin{bmatrix} Z_1 + Z_4 & -Z_4 & Z_4/N \\ -Z_4 & Z_4 + Z_1 & Z_4/N \\ N & N & -1 \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \end{bmatrix} = \begin{bmatrix} V_1 - V_4 - N\,V_3 \\ V_4 - V_2 - N\,V_3 \\ 0 \end{bmatrix} \tag{B.18}$$

In addition to isolation between ports, it is sometimes desired to have all 4 ports matched for maximum power transfer. This means that a conjugate match exists between each port and its terminating impedance. Consider the input impedance seen by the source that is connected to port 1. For a conjugate match at port 1, we require that $Z_{in1} = Z_1^*$. The impedance Z_{in1} is equal to the voltage at terminal 1, which is $V_1 - I_a Z_1$, divided by the current flowing into this terminal, which is I_a. Thus, $Z_{in1} = \frac{V_1}{I_a} - Z_1$. When solving for I_a we set V_2, V_3, $V_4 = 0$ in the equations B.28, because only port 1 is driven when calculating Z_{in1}. Solving for I_a:

$$I_a = \frac{\begin{vmatrix} V_1 & -Z_4 & Z_4/N \\ 0 & Z_4 + Z_1 & Z_4/N \\ 0 & N & -1 \end{vmatrix}}{D}$$

$$I_a = -(2Z_4 + Z_1)V_1/D \tag{B.19}$$

where

$$D = -(2Z_4 + Z_1)^2 \tag{B.20}$$

Thus

$$\frac{V_1}{I_a} = 2Z_4 + Z_1. \tag{B.21}$$

The input impedance is then

$$Z_{in1} = 2Z_4. \tag{B.22}$$

For a conjugate match at port 1 we need

$$Z_{in1} = {Z_1}^* \tag{B.23}$$

so

$$Z_4 = \frac{1}{2} Z_1^* \tag{B.24}$$

Notice that the combination of the biconjugate conditions (equations B.17) and equation B.24 determine the terminations at ports 2, 3, and 4 once the termination at port 1 is specified. There are no additional degrees of freedom left once biconjugacy and a conjugate match at one port are enforced. This implies that the input impedances at the other ports (2, 3, and 4) are determined once the conjugate match at port 1 is enforced. It is not hard to show that a biconjugate system with one port matched for maximum power transfer will automatically be conjugate matched at all 4 ports. How wonderful! A biconjugate transformer that is conjugately matched at all 4 ports is called a hybrid transformer. The terminating impedances for a hybrid transformer system must satisfy the following conditions:

$$\begin{aligned} Z_1 &= Z & \text{(B.25)} \\ Z_2 &= Z & \text{(B.26)} \\ Z_3 &= Z^*/2N^2 \\ Z_4 &= Z^*/2 \end{aligned}$$

If port 1 is terminated with a resistance, R. Then port 2 must also be terminated with R, port 3 must be terminated with $R/(2N^2)$ and port 4 must be terminated with $R/2$. Notice that if the turns ratio is chosen to be $N = \frac{1}{\sqrt{2}}$ then $Z_1 = Z_2 = Z_3 = R$ and $Z_4 = R/2$. We will assume that the port terminations are resistive and that $N = \frac{1}{\sqrt{2}}$ in the discussions to follow.

B.3 Applications of the Hybrid Transformer

B.3.1 Power Splitters

B.3.1.1 180-degree splitter

Suppose a signal is applied to port 3 as in Figure B.3.

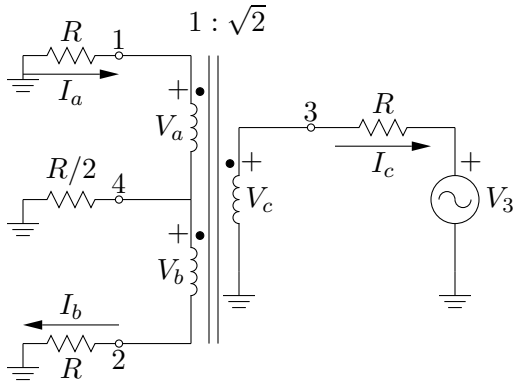


Figure B.3: A hybrid transformer driven at port 3.

The system of equations for this hybrid transformer is

$$\begin{bmatrix} \frac{3}{2}R & -\frac{1}{2}R & \frac{1}{\sqrt{2}}R \\ -\frac{1}{2}R & \frac{3}{2}R & \frac{1}{\sqrt{2}}R \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -1 \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{2}}V_3 \\ -\frac{1}{\sqrt{2}}V_3 \\ 0 \end{bmatrix} \tag{B.27}$$

Since port 4 is conjugate to port 3, we already know that there is no response at port 4. Hence, the current leaving port 4 must be zero, which implies that $I_b = I_a$. Let's solve for I_a:

$$I_a = \frac{\begin{vmatrix} -\frac{1}{\sqrt{2}}V_3 & -\frac{1}{2}R & \frac{1}{\sqrt{2}}R \\ -\frac{1}{\sqrt{2}}V_3 & \frac{3}{2}R & \frac{1}{\sqrt{2}}R \\ 0 & \frac{1}{\sqrt{2}} & -1 \end{vmatrix}}{D}.$$

The solution is

$$I_a = I_b = -V_3\frac{\sqrt{2}}{4R}.$$

Denoting the voltages at terminals 1 and 2 by V_1 and V_2, respectively, we have $V_1 = -I_aR$ and $V_2 = I_bR$, or

$$V_1 = \frac{\sqrt{2}}{4}V_3, \quad V_2 = -\frac{\sqrt{2}}{4}V_3.$$

Hence, the voltages at ports 1 and 2 have equal amplitudes, and are 180 degrees out of phase. The power delivered to terminations 1 and 2 is

$$P_1 = P_2 = \frac{1}{2}\frac{|V_1|^2}{R} = \frac{1}{16}\frac{|V_3|^2}{R}.$$

Note that P_1 and P_2 are each exactly one half of the power available from the source connected to port 3. This configuration of the hybrid transformer is called the 180-degree power splitter.

B.3.1.2 In-phase splitter

Suppose a signal is applied to port 4 as in Figure B.4. The system of equations for this hybrid transformer is

$$\begin{bmatrix} \frac{3}{2}R & -\frac{1}{2}R & \frac{1}{\sqrt{2}}R \\ -\frac{1}{2}R & \frac{3}{2}R & \frac{1}{\sqrt{2}}R \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -1 \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \end{bmatrix} = \begin{bmatrix} -V_4 \\ V_4 \\ 0 \end{bmatrix} \tag{B.28}$$

Since port 4 is conjugate to port 3, we already know $I_c = 0$. Let's solve for I_a:

$$I_a = \frac{\begin{vmatrix} -V_4 & -\frac{1}{2}R & \frac{1}{\sqrt{2}}R \\ V_4 & \frac{3}{2}R & \frac{1}{\sqrt{2}}R \\ 0 & \frac{1}{\sqrt{2}} & -1 \end{vmatrix}}{D}.$$

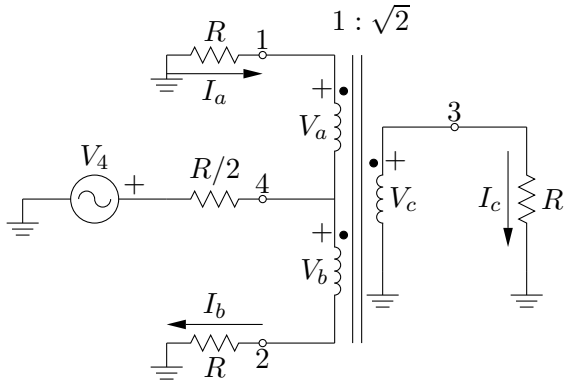


Figure B.4: A hybrid transformer driven at port 4.

The solution is

$$I_a = -V_4 \frac{1}{2R}.$$

Similarly,

$$I_b = +V_4 \frac{1}{2R}.$$

The voltages at terminals 1 and 2 are $V_1 = -I_a R$ and $V_2 = I_b R$, or

$$V_1 = \frac{1}{2} V_4, \quad V_2 = \frac{1}{2} V_4.$$

Hence, the voltages at ports 1 and 2 have equal amplitudes, and are in phase. The power delivered to terminations 1 and 2 is

$$P_1 = P_2 = \frac{1}{2} \frac{|V_1|^2}{R} = \frac{1}{8} \frac{|V_4|^2}{R}.$$

Note that P_1 and P_2 are each exactly one half of the power available from the source connected to port 4. This configuration of the hybrid transformer is called the in-phase power splitter.

B.3.2 Sum or Difference Combiners using a Hybrid Transformer

The hybrid transformer has the property that the phase shift between three of the four ports is zero and the phase shift to the remaining port will be 180°. This property can be used to realize both adding and subtracting signal combiners. For example, consider the response to applied signals V_1 and V_2 as shown in Figure B.5. [1]The output at port 3 will

[1]Note that since ports 1 and 2 are conjugate, the signal generators connected to ports 1 and 2 are isolated — this means that the signal from generator 1 will not appear at the output terminals of generator 2 and vice versa. Without this isolation, the signal from generator 1 could mix with the signal from generator 2 in the output stage of either generator to produce unwanted intermodulation products.

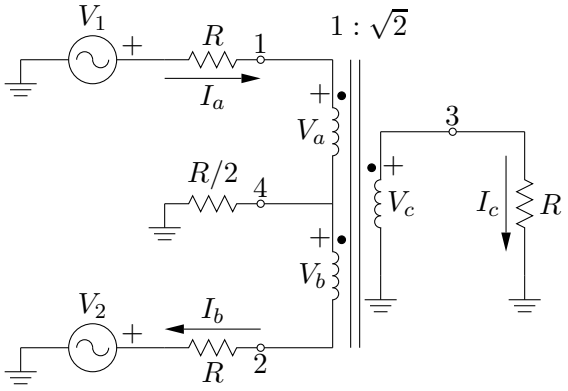


Figure B.5: Hybrid transformer used to combine two signals.

be proportional to I_c where

$$I_c = \frac{D_c}{D} = \frac{\begin{vmatrix} \frac{3}{2}R & -\frac{1}{2}R & V_1 \\ -\frac{1}{2}R & \frac{3}{2}R & -V_2 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{vmatrix}}{D}.$$

The solution is

$$I_c = \frac{\sqrt{2}}{4R}(V_1 - V_2)$$

The voltage at port 3 is $V_3 = I_c R$, so

$$V_3 = \frac{\sqrt{2}}{4}(V_1 - V_2).$$

The outport at port 3 is proportional to the difference between the applied signals.

The output at port 4 will be proportional to $I_a - I_b$. Using the fact that ports 1 and 2 are conjugate, and noting that generators 1 and 2 each see a matched impedance, R, it is obvious that $I_a = \frac{V_1}{2R}$ and $I_b = -\frac{V_2}{2R}$. Hence

$$V_4 = (I_a - I_b)\frac{R}{2} = \frac{1}{4}(V_1 + V_2).$$

The output at port 4 is proportional to the sum of the applied signals.

B.4 References

1. Smith, Jack, *Modern Communications Circuits*, Second Edition, WCB/McGraw-Hill, 1998.

2. Sartori, Eugene P., "Hybrid Transformers", IEEE Transactions on Parts, Materials and Packaging, Vol. PMP-4, No. 3, September 1968, pp 59-66.

Appendix C

Useful Constants and Trigonometric Identities

speed of light in free-space	$c = 2.998 \times 10^{8}\ m\ s^{-1}$
permittivity of free-space	$\epsilon_0 = 8.854 \times 10^{-12}\ F\ m^{-1}$
permeability of free-space	$\mu_0 = 4\pi \times 10^{-7}\ H\ m^{-1}$
Boltzmann constant	$k = 1.381 \times 10^{-23}\ J\ K^{-1}$
elementary charge	$q = 1.602 \times 10^{-19}\ C$
Planck constant	$h = 6.626 \times 10^{-34}\ J\ s$

Table C.1: Some useful constants

$$\sin a = \frac{e^{ja} - e^{-ja}}{2j}$$
$$\cos a = \frac{e^{ja} + e^{-ja}}{2}$$
$$\tan a = \frac{\sin a}{\cos a}$$
$$\sec a = \frac{1}{\cos a}$$
$$\csc a = \frac{1}{\sin a}$$
$$\sin(a \pm b) = \sin a \cos b \pm \cos a \sin b$$
$$\cos(a \pm b) = \cos a \cos b \mp \sin a \sin b$$
$$\tan(a \pm b) = \frac{\tan a \pm \tan b}{1 \mp \tan a \tan b}$$
$$\sin a \sin b = \frac{1}{2}\cos(a-b) - \frac{1}{2}\cos(a+b)$$
$$\sin a \cos b = \frac{1}{2}\sin(a+b) + \frac{1}{2}\sin(a-b)$$
$$\cos a \cos b = \frac{1}{2}\cos(a+b) + \frac{1}{2}\cos(a-b)$$
$$\cos^2 a = \frac{1}{2}(1 + \cos(2a))$$
$$\sin^2 a = \frac{1}{2}(1 - \cos(2a))$$
$$\cos^3 a = \frac{3}{4}\cos a + \frac{1}{4}\cos(3a)$$
$$\sin^3 a = \frac{3}{4}\sin a - \frac{1}{4}\sin(3a)$$
$$\sinh a = \frac{e^{a} - e^{-a}}{2}$$
$$\cosh a = \frac{e^{a} + e^{-a}}{2}$$
$$\tanh a = \frac{\sinh a}{\cosh a} = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}}$$

Table C.2: Some trigonometric identities.