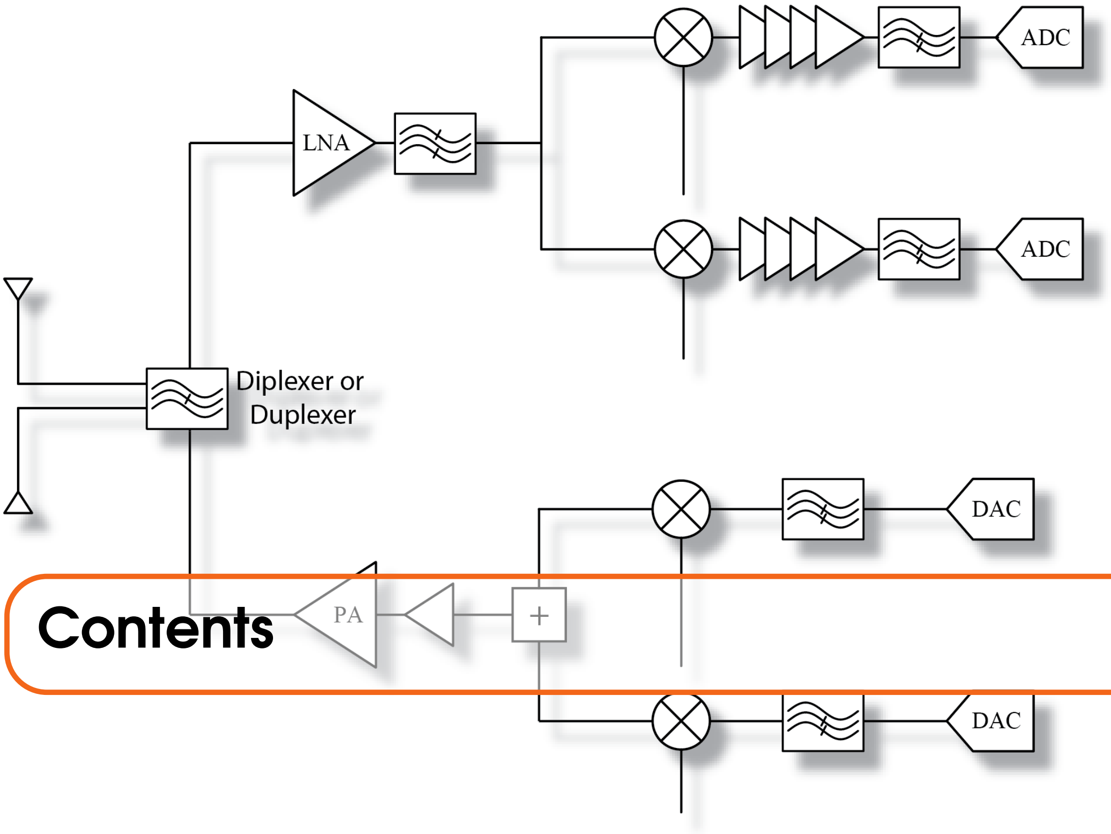
Communication Circuits

RF, Microwaves, mm-Waves, THz, and Beyond

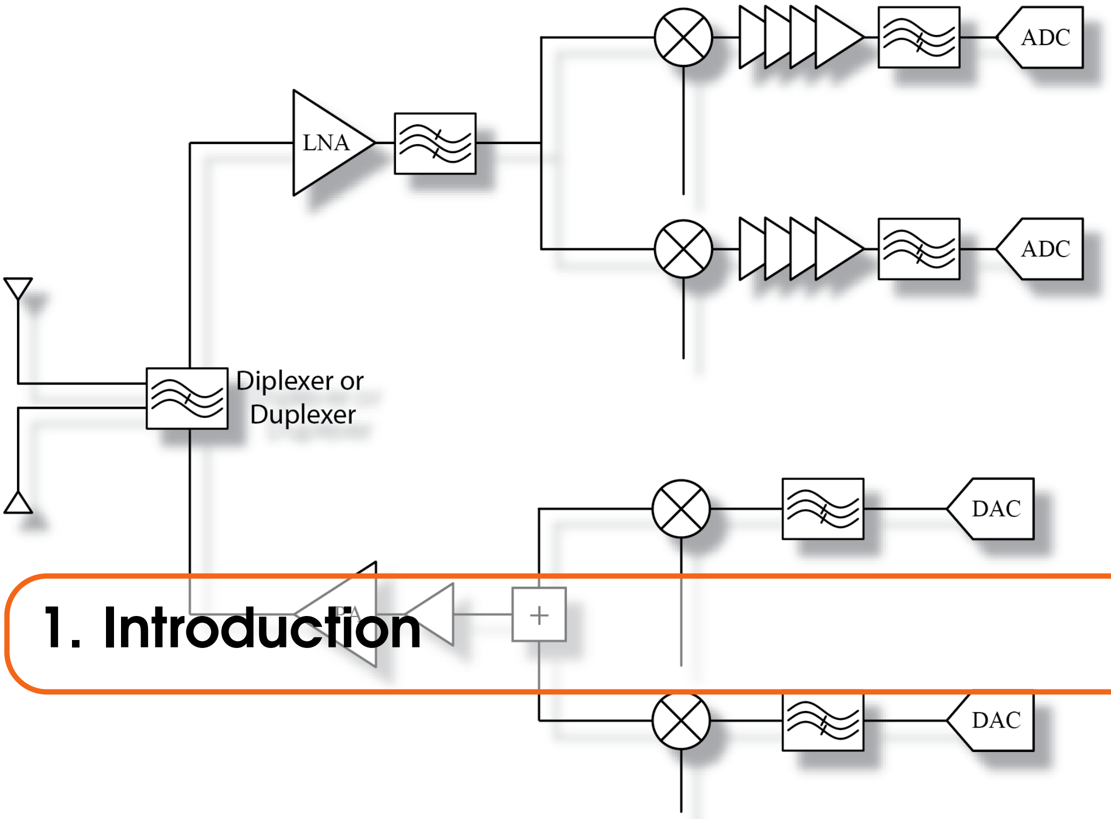
Ali M. Niknejad

PUBLISHED BY ALI NIKNEJAD

RFIC.EECS.BERKELEY.EDU/RFDOG.HTML



Contents



1. Introduction

The goal of every communication system is to transmit and receive a signal through a channel (wireless or wired) in the presence of noise and interference. As signals propagate, they are attenuated and corrupted by noise. Noise is both environmental noise, such as thermal noise, which is due to random motion of charged particles, and also includes “noise" generated by other users of the channel. Also, surprisingly, noise and interference can be self-inflicted, due to multi-path propagation or unwanted leakage paths, covered in Section 1.9.2.

1.1 Wireless and Wired Communication

Most information sources are baseband (BB) in nature, where we arbitrarily define the bandwidth *BW* as the highest frequency of interest. This usually means that beyond the *BW* the integrated energy is negligible compared to the energy within the bandwidth. For example, the bandwidth of some common signals include:

- High fidelity audio: 20 kHz
- Analog telephone quality audio: 5 kHz
- Uncompressed analog video: ∼10 MHz
- Data traffic from 802.11 b/g WLAN: 22 MHz
- Data traffic from 802.11 n/a/ax: 20/40/80/160 MHz
- HD Video (HDMI 1.3+): ∼340 MHz

The source is often compressed to conserve bandwidth. Lossless compression (LZW like Zip files) or lossy (like MP3, JPEG, or MPEG video) can be used depending on the application. Compressing the financial information in your bank account versus the image of a cat, the choice is usually clear.

What makes communication challenging, and for the engineer interesting, is the fact that the “channels" are often physically long or wireless in nature. Long wires or cables underground or across the ocean floor are important examples, but even the cable connecting your PC to the monitor is long enough such that extra care is required to transfer the information without excessive distortion or noise. Let’s look at the various impairments we encounter in communication systems.

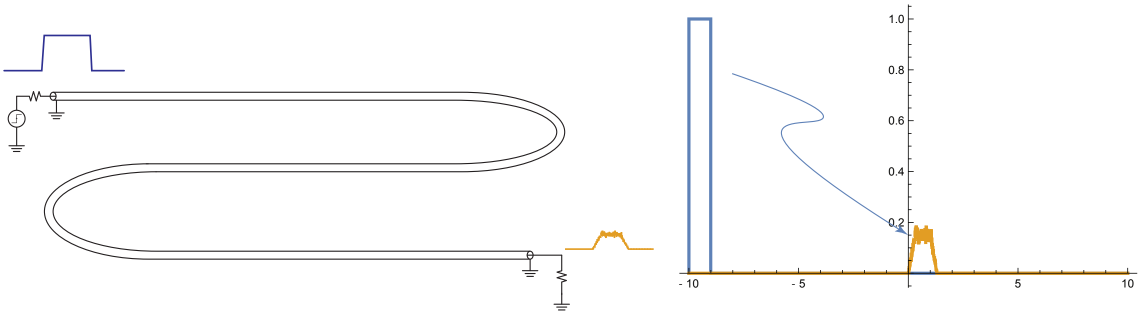


Figure 1.1: In an ideal communication system, the received pulse is a delayed and possibly scaled copy of the transmitted pulse. In practice, the received pulse is accompanied by noise, dispersion, and reflections.

1.1.1 Signal Attenuation / Channel Bandwidth

When sending high speed data through a cable, we have to deal with several non-idealities such as *attenuation*, *dispersion*, *noise*, *distortion* and *reflections*, as shown in Fig. 1.1. All of these terms have technical distinct meanings, but for now let's just think of all of these as sources of impairment that cause the received pulse shape to differ from the transmitter pulse, both in height, width and even content.

Waveform Dispersion and Inter-Symbol Interference (ISI)

A perfectly "linear" ideal channel must preserve the pulse shape as the signal travels through the line. We generally tolerate attenuation due to loss if the pulse shape is largely unchanged from source to load, since signal amplification can restore the pulse size. The pulse shape, especially pulse widening due to dispersion, is problematic as it limits the data rate. For a linear channel note that we must not only be concerned with the amplitude frequency response, but also the phase response. A delayed version of the pulse at the receiver only experiencing attenuation α is described by:

$$s_r(t) = \alpha \cdot s_t(t-\tau) \tag{1.1}$$

or in the frequency domain:

$$S_r(\omega) = \alpha \cdot S_t(\omega) e^{j\omega\tau} \tag{1.2}$$

where τ is the delay time. Thus the amplitude response is flat with frequency:

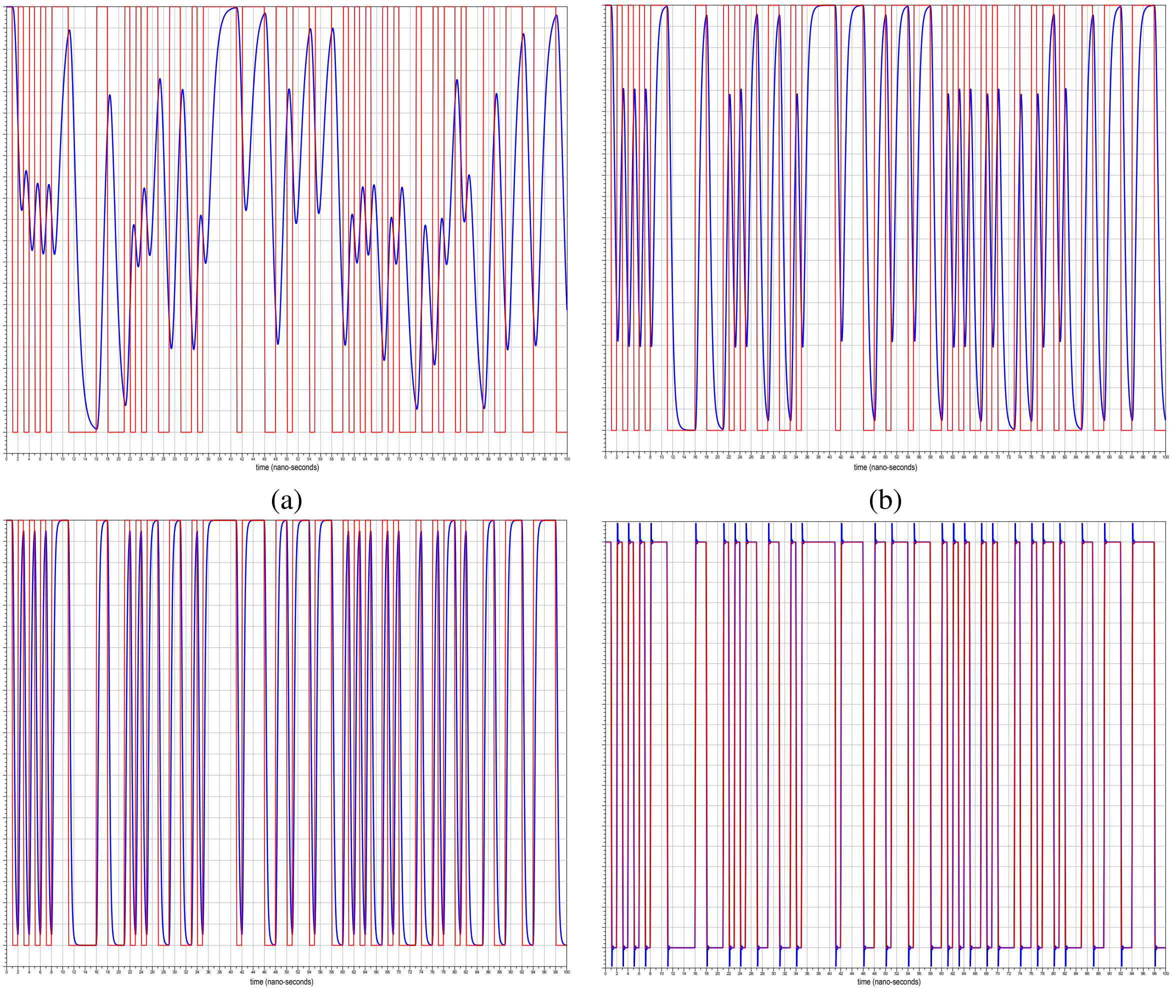
$$|H(\omega)| = \left| \frac{S_r(\omega)}{S_t(\omega)} \right| = \alpha \tag{1.3}$$

and the phase response is linear

$$\angle H(f) = \omega\tau \tag{1.4}$$

which implies constant group delay, which really means that all frequencies arrive in the correct phase at the destination in a manner to reconstruct the original signal.

In practice the attenuation is often frequency selective, with higher frequency bands more heavily attenuated than lower bands, in other words the channel is low-pass in nature. Some channels, such as waveguides, are bandpass and won't propagate below a cut-off frequency. Frequency dependent attenuation leads to distortion in the transmitted waveform. Imagine sending



(a)

(b)

(c)

(d)

Figure 1.2: The input (red) and output (blue) pulses traveling through a low-pass channel. The data clock is at 1 GHz and the channel bandwidth is (a) 0.25 GHz, (b) 0.5 GHz, (c) 1 GHz, and (d) 10 GHz. Only the 10 GHz channel preserves the pulse edges (with some ringing) whereas below 1 GHz bandwidth the pulses are eigher too distorted or experiencing incomplete settling.

a suit to the washer and getting back pants scaled by a factor of 90% but the jacket scaled by 80%. If the scaling were uniform, you could still potentially wear the suit if you went on a diet, but with non-uniform scaling, the suit is just not going to work anymore as a complete unit. In the same fashion, if you send a pulse through a transmission line and it's low-pass filtered, then the "edges" of the waveform will be softened and the pulse is no longer an identical copy of the transmitted pulse, as shown in Fig. 1.2. If the frequency response is too narrow, then there's incomplete settling and the pulse height will be attenuated.

Even if the information is binary in nature, the presence of softer edges results in pulse attenuation and limits the maximum data rate that we can transmit/receive information. Generally the bandwidth of the channel needs to be at least as large as clock rate, but often we prefer to preserve the waveform edges, or the rise/fall time of the pulse, which demands much higher bandwidth proportional to $1/t_{rise}$. Usually this is the case for short range digital links. Higher range links must compromise pulse shape to preserve bandwidth, and usually specially engineered pulse shapes are used, which also helps to avoid other problems, such inter-symbol interference,

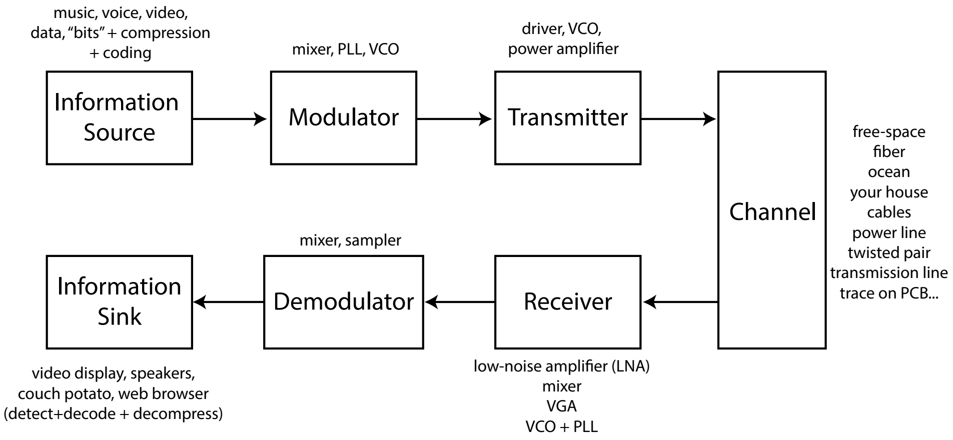


Figure 1.3: A block diagram representation of a communication system.

described next.

Inter Symbol Interference (ISI) is the process where one bit interferences with adjacent bits due to incomplete settling time or due to multiple pulses arriving at the load with different gain/delay profiles. This can happen if the channel bandwidth is too low, as shown in Fig. 1.2, but it may also occur due to reflections or multi-path, where a pulse takes more than one path from the source to the destination or several delayed copies arrive at the destination. As we'll learn in Ch. **??**, when impedances are not well matched, signals reflect and cause ISI. Equalization is used at the source and receiver to compensate for the non-ideality of the channel. For example, the higher frequencies can be pre-emphasized using a high-pass channel. The phase response can be adjusted using an all-pass filter to overcome the non-constant group delay. More sophisticated equalizations can intentionally introduce ISI to cancel the channel ISI, making the channel response approach a single pulse rather than multiple copies. But in all of these situations the "channel" has to be characterized first. In other words, we need to send training data to discover the channel response.

1.1.2 Wireless Propagation in Free-Space

LOS emphasis Friis

1.1.3 Multi-Path Propagation

Multi-path propagation refers to the fact that there's more than one path from source to destination, and delayed copies of the signal appear like noise.

1.2 Antennas for Wireless

1.2.1 Radiation Pattern

1.2.2 Antenna Directivity

1.2.3 Arrays of Antennas

Phased Arrays

1.3 Block Diagram of Communication System

A typical communication system, shown in Fig. 1.3, can be partitioned into a transmitter, a channel, and a receiver. In this book we will study the circuits that interface to the channel from the receiver/transmitter. These circuits are at the "front-end" of the transceiver and operate at high frequency. The baseband circuits are responsible for encoding, compression, modulation,

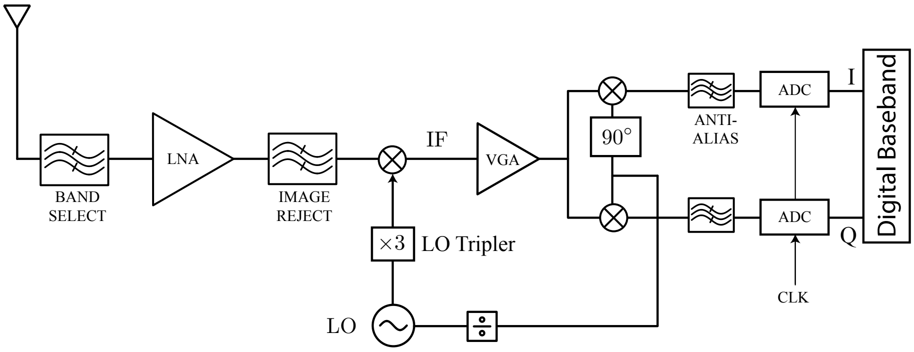


Figure 1.4: The ubiquitous super-heterodyne receiver architecture.

demodulation, and detection. These circuits are typically implemented in mixed-signal and/or digital form. The focus of this book is the radio frequency (“RF") circuitry and the baseband analog circuitry, everything preceding the analog-to-digital converter. Although the term “RF" is a generic term, sometimes people call everything below a certain frequency (say 1 GHz in the old days) RF, and everything above it microwaves. Today the distinction is less useful as the same technology and circuit techniques can be used to realize circuits from essentially audio bands all the way to 300 GHz or more, the so-called mm-wave and sub-THz bands. In this book we will use the term RF to encompass all of these frequency bands.

1.3.1 Transmitter Block Diagram

1.3.2 Simple “AM" Receiver

1.3.3 Simple “FM" Transmitter/Receiver

1.3.4 The Modern Receiver

The super-heterodyne receiver is shown in Fig. 1.4. Don’t worry about what we mean by “super-heterodyne," we’ll cover that later. For now, let’s focus on the important *active* and *passive* building blocks in this system. Passive blocks include switches, filters, and resonators. Active building blocks draw DC current and consume power whereas passive components such as filters or resonators can be realized without a power supply. Active blocks include:

- LNA: Low noise amplifier used to amplify (gain up) the signal
- MIXER/LO: The mixer (multiplier) and “Local” Oscillator (LO) is used to frequency translate the signal from RF to baseband. Think of this block as a multiplier we need to modulate and demodulate the signal in the AM modulator we described above.
- VGA: Variable Gain Amplifier (or PGA for programmable gain amplifier) is used to provide sufficient amplification to detect the signal. Due to a widely varying signal levels, the gain is programmable.
- ADC: Analog to Digital Converter simply converts the data from analog format into digital format. It discretizes both the time by sampling and the amplitude by quantization.
- DSP: Digital Signal Processor

1.3.5 The Modern Transmitter

A so-called heterodyne transmitter block diagram is shown in Fig. 1.5. As before, do not worry about the word “heterodyne" as we will discuss this in detail in later chapters. For now, let’s focus on the individual blocks in the transmitter. In addition to passive filters and switches, we have the following important active building blocks include:

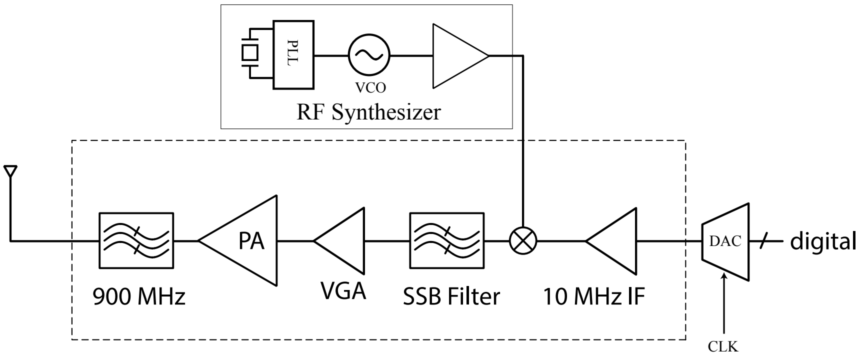


Figure 1.5: Block diagram of a heterdyne transmitter.

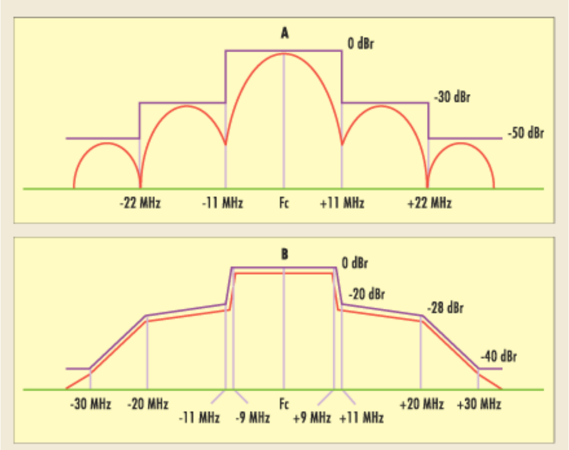


Fig.1 The spectral mask for 802.11b (a) is defined as -30 dBr beyond +/– 11 MHz and –50 dBr beyond 22 MHz. For OFDM, it is 0 dBr to 9 MHz, –20 dBr at 11 MHz offset, –28 dBr at 20 MHz, and –40 dBr beyond 30 MHz.

Figure 1.6: The spectral mask dictates the close-in transmission characteristics of a transmitter.

- DAC: Digital to Analog Converter takes digital bits and converts them into an analog waveform.
- Mixer: Up-conversion mixer translates the data from baseband (near DC) to RF, essentially performing modulation for simple AM systems.
- VGA: To select desired output power.
- Frequency Synthesizer: Synthesize carrier frequency.
- PA: Power Amplifier, used to drive sufficient power into the antenna.

Transmitter Spectrum

The transmitter must amplify the modulated signal and deliver it to the antenna (or cable, fiber, etc) for transmission over the communication medium. Generating sufficient power in an efficient manner for transmission is a challenging task and requires a carefully designed power amplifier. Even the best RF power amplifiers do this with only about 60% peak efficiency at gigahertz frequencies. In practice the average efficiency is much lower because we will learn that the efficiency depends on the output power and generally drops with lower power levels. If the envelope of the signal varies, the efficiency also varies. The transmitted spectrum is also corrupted by phase noise and distortion (see chapters **??** and **??**). This corrupts the signal modulation (loss in Error Vector Magnitude (EVM)), causes adjacent channel interference (spectral regrowth), and generates harmonics into other bands. Communication standards specify a spectral mask (Fig. 1.6) that must be satisfied to avoid interference, limiting the power in adjacent bands. FCC limits dictate that power in harmonics must also be sufficiently low.

1.4 Modulation and Demodulation

In most wireless communication systems, the information signal $s(t)$ is modulated onto the amplitude and/or the phase of a carrier signal. The carrier signal is a sinusoidal signal, so the transmitted signal $T(t)$ can be written as

$$T(t) = A[s(t)]\cos(\omega_c t + \Phi[s(t)]) \tag{1.5}$$

where $A[s(t)]$ is the amplitude modulation and $\Phi[s(t)]$ is the phase modulation (possibly non-linear functions). Modulation has many advantages over transmitting the baseband waveforms directly. First, it allows multiple users to share the spectrum by choosing different carrier frequencies ω_c. (due to the orthogonality of sinusoidal signals). By appropriately filtering, we can always select the desired band of interest, provided the modulated signals do not overlap in the frequency domain. For example, suppose we simply amplitude modulate (AM) a signal such that

$$T(t) = s(t)\cos(\omega_c t) \tag{1.6}$$

In this case the function $A[s(t)] = 1$, a linear modulation scheme. In this AM scheme the modulator is simply a multiplier that multiplies the waveform $s(t)$ with a “Local Oscillator" or LO signal $\cos(\omega_c t)$. After we receive the above signal, we can first filter out all other signals in the band and then de-modulate the received sginal by multiplying with the carrier once more[1]

$$R(t) = \alpha \cdot s(t-\tau)\cos(\omega_c t + \phi) \times \cos(\omega_c t) \tag{1.7}$$

α represents the signal attenuation as it travels from the source to the destination, and ϕ is the phase shift, $\phi = \omega_c \tau$ due to the flight delay τ. Here we assume that only one copy travels from the source to the destination, so there is negligible multi-path. [2] Expanding the first term in $\cos(\omega_c t + \phi)$ in $R(t)$,

$$\cos(\omega_c t + \phi) = \cos(\omega_c t)\cos(\phi) - \sin(\omega_c t)\sin(\phi) \tag{1.9}$$

we can write $R(t)$ as

$$R(t) = \alpha s(t-\tau)\left(\cos(\omega_c t)^2\cos(\phi) - \cos(\omega_c t)\sin(\omega_c t)\sin(\phi)\right) \tag{1.10}$$

Substituting $\cos^2(x) = \frac{1}{2}(1+\cos 2x)$ and $\cos x \sin x = \frac{1}{2}\sin 2x$, we have

$$R(t) = \alpha s(t-\tau)\frac{1}{2}\left((1+\cos(2\omega_c t))\cos\phi - \sin(2\omega_c t)\sin\phi\right) \tag{1.11}$$

By simply low-pass filtering $R(t)$, we can extract the original AM signal (delayed)

$$R'(t) = LPF[R(t)] \approx \alpha s(t-\tau)\frac{1}{2} \tag{1.12}$$

[1] The sharp reader may inquire how we generate another copy of the same sinusoid, at the same frequency and phase as the transmitter.

[2] ϕ is the phase shift that the signal experiences. Note that we can also calculate this phase shift by relating it to the distance traveled by $\phi = kx$, where $k = 2\pi/\lambda$ is the wave number. Since $\lambda f = c$, the propagation velocity c, and the propagation delay is simply $\tau = x/c$, we can write ϕ as

$$\phi = kx = \frac{2\pi}{\lambda}x = \frac{2\pi f x}{c} = 2\pi f \tau = \omega_c \tau \tag{1.8}$$

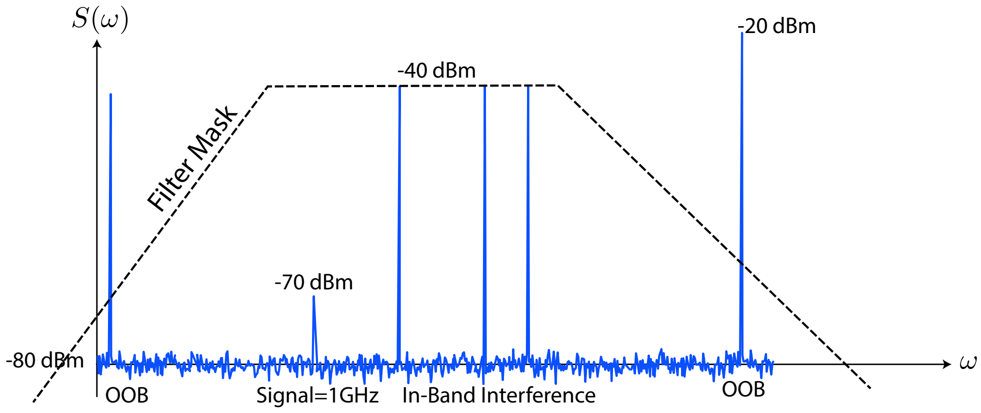


Figure 1.7: The spectrum seen by a receiver antenna contains the desired signal accompanied by many blockers, both out-of-band, which can be filtered out, and also in-band, or other users of the spectrum. Due to the unavailability of low cost narrow and tunable filters, all of these in-band signals enter the receiver.

This is the basis of most wireless communication systems. We will show later that phase modulation can actually be treated as an amplitude modulation if we use two orthogonal carriers, sin and cos, sometimes referred to as complex modulation. We can also convert phase modulation into amplitude modulation by taking the derivative of a signal.

1.5 Goal of a Communication System

1.5.1 Receiver Spectrum: Finding a Needle in a Haystack

The received signal is often accompanied by interference, arising from other users of the same band or users in other bands – known as Out-of-Band (OOB) blockers. The receiver must be able to listen to a weak channel when others may be significantly larger. The typical spectrum at the input of the receiver is shown in Fig. 1.7. Note that the desired signal is barely larger than the noise, and may be even much weaker than other signals. In addition to out-of-band interfering signals, which can be easily filtered out, the receiver must also contend with strong in-band interferers. These nearby signals are often other channels in the spectrum, or other users of the spectrum.

The dynamic range (DR) is the ratio of the highest power a receiver can process relative to the lowest power, which is usually noise limited. The ratio in linear scale is often expressed as the difference in dB scale. The signal strength varies a great deal as the user moves closer or further from a base-station (access point). Due to multi-path propagation and shadowing, the signal strength varies in a time varying fashion. The *dynamic range* of a wireless signal is therefore very large, on the order of 60-100 dB.

1.5.2 Don't Throw Out the Baby with the Bathwater!

The worse case scenario is the so called near-far problem (Fig. 4.22), in which you're trying to receive a distant weak signal, but you're also“hearing" a nearby strong signal (jammer or blocker). In other words you're trying to understand a whisper in the presence of a shout. Sometimes, in full duplex systems, the jammer is your own transmitter. In frequency division duplexing, the transmitter is on at the same time but with a slight offset in frequency. We require low noise and high linearity to successfully converse in this scenario.

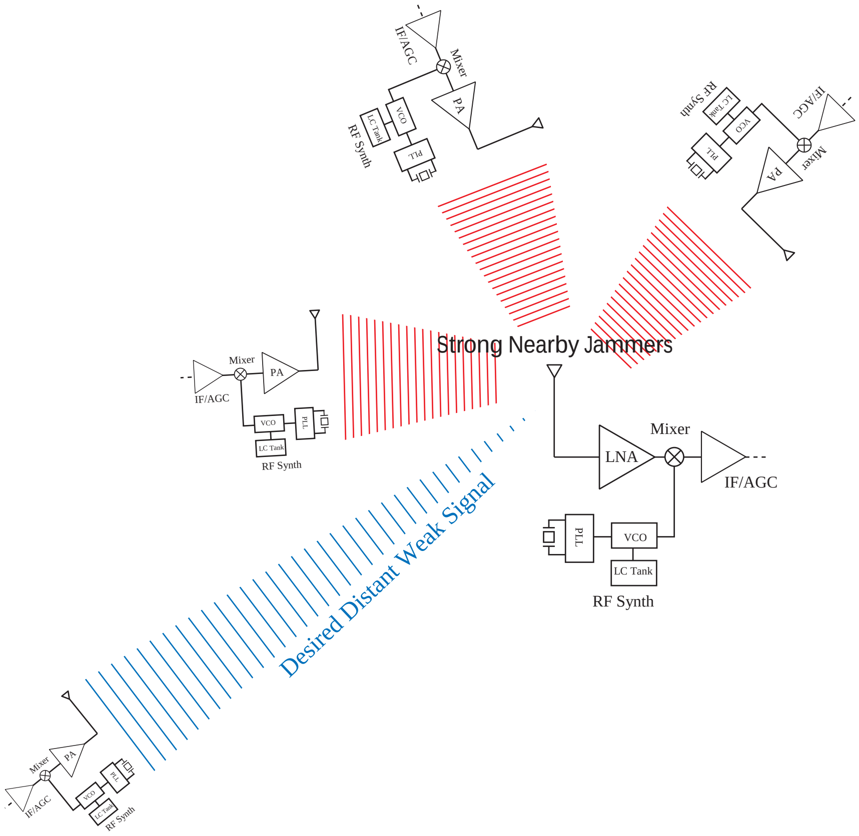


Figure 1.8: The so-called “near-far" problem, when a nearby transmitter is very loud and we’re trying to detect a distant weak signal.

1.5.3 Goals: Amplification

The power in communication systems is often measured in the dBm scale, or the log power measured relative to a 1 mW reference. E.g. a power level of 10 mW can be expressed as 10 dBm

$$10 \cdot \log\left(\frac{10\,\mathrm{mW}}{1\,\mathrm{mW}}\right) = 10\,\mathrm{dBm} \tag{1.13}$$

On your laptop or cellular phone, you can often see the signal strength expressed in dBm units.[3]

Amplification of weak signals is a major goal of a communication system. As we’ll learn later in this course, amplifier comes with the cost of additional noise and distortion imparted on the signal. This is problematic if the signals of interest are only marginally larger than the intrinsic noise produced by the system. Additionally, high gain for the interference signals can easily “rail” or saturate our amplifiers unless we carefully filter them out.

As an example, say your WiFi (also known as WLAN) on your laptop is receiving a signal with strength $-70\,\mathrm{dBm}$. This corresponds to a power of $P = 10^{-10}\,\mathrm{W} = 100\,\mathrm{pW}$. The voltage on the antenna can be approximated by

$$P = \frac{V_{ant}^2}{2Z_0} \tag{1.14}$$

[3] A handy utility on the Mac OS X operating system is the “Airport" command line utility (no longer distributed by default) that lists all the available WiFi channels, the power levels, and the current transmission rate and modulation format.

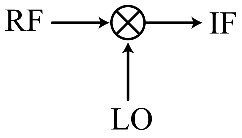


Figure 1.9: A mixer is a block that effectively multiplies a locally generated oscillation (LO) with an input signal, allowing us to move the spectrum of the signal up or down in frequency.

where $Z_0 = 50\Omega$, is the assumed antenna impedance. Solving for V_{ant}

$$V_{ant} = \sqrt{2Z_0P} = \sqrt{2 \cdot 50 \cdot 10^{-10}} = 10^{-4}\,\text{V} = 100\,\mu\text{V} \tag{1.15}$$

This is a very small signal, but in RF terms it's a “healthy" signal since it does not come anywhere close to the noisefloor of the system. Let's consider an even smaller signal next.

1.5.4 Goals: Filtering

A cell/mobile phone is designed to work with very weak signals. For instance for $P = -100\,\text{dBm}$, or $P = 10^{-13}\,\text{W}$, we have

$$V = \sqrt{100 \cdot 10^{-13}} = \sqrt{10} \times 10^{-6} \sim 3\,\mu\text{V} \tag{1.16}$$

This is indeed a tiny signal and it might be close to the noise floor of the system. We need a voltage gain of about 10^5 or 50 dB to bring this signal into the range for baseband processing (300mV). Now imagine an interference signal of strength $-40\,\text{dBm}$, or about 3 mV. This may seem like a small signal, but it effectively limits the gain of our system to about 300, assuming a 1V supply. Unless we employ a very high resolution ADC (expensive, bulky, power hungry), we must filter out this interference before digitization. To see this, consider that without filtering we can only raise our desired signal to a voltage level of $300 \times 3\,\mu\text{V} = 900\,\mu\text{V}$. This must be stronger than the quantization noise floor of the ADC. If we assume a factor of 10, the quantization noise of the ADC must be $90\,\mu\text{V}$, requiring a resolution of

$$\log_2\left(\frac{1\,\text{V}}{90\,\mu\text{V}}\right) = 13.4\,\text{bits} \tag{1.17}$$

sampled at twice the bandwidth of the signal.

1.5.5 Goals: Frequency Translation

Frequency translation is a non-linear or time-varying operation, since an LTI system cannot perform frequency translation. In most scenarios, this function is performed with a mixer, shown in Fig. 1.9. As discussed previously, the received signal can be represented in the following form

$$V_r = A(t)\cos\left(\omega_c t + \phi(t)\right) \tag{1.18}$$

The frequency ω_c is the “carrier frequency” since the modulation rides on top of this signal. The term $A(t)$ represents the amplitude modulation, or AM. The term $\phi(t)$ is the phase modulation (PM). Frequency modulation, or FM, can also be achieved through $\phi(t)$

$$A\cos(\omega_c t + \delta\omega \int_0^t m(t)dt) \tag{1.19}$$

where $m(t)$ is the normalized modulation waveform and $\delta\omega$ is the maximum frequency deviation.

E.g. in broadcast television we use AM for the video and FM for audio. A digital modulation scheme may involve AM, FM, PM, or some combination.

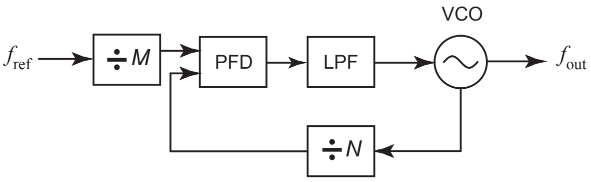


Figure 1.10: A phase-locked loop (PLL) is usually the way a frequency f_{out} is synthesized from a stable reference source, f_{ref}, often derived from a crystal oscillator.

1.5.6 Goals: Frequency Synthesis

Since carrier frequencies are used for RF modulation, a transmitter and receiver need to synthesize a precise and stable reference frequency. Since the reference frequency changes based on which "channel" is employed, the synthesizer must be tunable. Think of the tuning "knob" on a radio receiver. The reference signal is generated by a voltage-controlled oscillator (VCO) and "locked" to a much more stable reference signal, usually provided by a precision quartz crystal resonator (XTAL). A phase-locked loop (PLL) synthesizer is a feedback system employed to provide the locking and tuning. It employs a frequency divider, a phase/frequency detector, and filtering to achieve this goal.

1.6 Multiple Access

1.6.1 UWB vs. Narrowband Signaling

Narrowband has been favored since spectrum can be chopped up into channels and interference is easily managed. Ultra-wideband (UWB) uses short pulses or windowed carriers and thus occupies a very large bandwidth. Energy is spread across a wide bandwidth so transmit power has to be limited to avoid interference. In the U.S., UWB transmission in the 3-10 GHz frequency range is allowed if the radiated power meets part 15 requirements (15.5.F), about -41.3 dBm per MHz of bandwidth. Other bands are more restrictive in power transmission. Generating a UWB signal is relatively easy by "gating" an oscillator for a short duration. One can modulate the phase, amplitude, or position in time to impart information. Receiving a UWB signal is challenging due to the wideband nature of the signal and the weak transmission power.

1.6.2 Spectrum Regulation

1.6.3 Spectrum Sharing

Suppose we wish to transmit a discrete-time signal $s(t)$ through the channel as shown. Generally a signal cannot be transmitted directly since many users wish to share the spectrum. In this section we will present various schemes for sharing the spectrum, including Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA) or Code Division Multiple Access (CDMA). TDMA means that users take turns to transmit so that they don't corrupt each other's transmissions, which requires time synchronization. Without synchronization, user's will collide and this will require retransmissions. One solution is to use different bands, or FMDA. Let's look at this problem a bit more generally by thinking of the carrier signal as a code rather than simply a pure sine wave.

With this framework in mind, in order to share the channel, users need to multiply their signals by (preferably orthogonal) codes so that the receiver can easily tell one users transmission apart from other users. For example transmitter $t_1(t)$ and $t_2(t)$ can each send their signals as

$$t_1(t) = s_1(t)c_1(t) \tag{1.20}$$

$$t_2(t) = s_2(t)c_2(t) \tag{1.21}$$

This process is similar to modulation, since we modulate the signal $s(t)$ with a given code $c(t)$. These codes should have two properties: (1) Mutual orthogonality, and (2) Delta-function like autocorrelation. Property (1) means that if we take the dot product between these codes, practically realized by multiplying and averaging over the duration of the code

$$< c_1, c_2 >= \int_T c_1(t)c_2(t)dt \approx 0 \tag{1.22}$$

for any delayed versions of c_1 and c_2. Strong autocorrelation means

$$< c_1(t), c_1(t-\tau) >\approx 0 \tag{1.23}$$

for any delay τ. In essence, we can imagine mapping each symbol of $s(t)$ onto a code. In other words, the code is repeated and modulated by the amplitude of $s(t)$ during each symbol time interval. How do we demodulate this signal? For now, let's ignore any echoes in the channel and consider receiving the composite signal

$$r(t) = \alpha s_1(t-\tau_1)c_1(t-\tau_1) + \beta s_2(t-\tau_2)c_2(t-\tau_2) \tag{1.24}$$

where α is the attenuation experienced by the first signal and β is likewise the attenuation of the second signal. Note that since we receive delayed copies of the signals, if we wish to receive signal s_1 in the presence of s_2, we need to know the delay τ_1. For now, suppose this information is known. Then we take the dot product of $r(t)$ with $c_1(t)$

$$< r(t), c_1(t-\tau_1) >= \alpha s_1(t-\tau_1) < c_1(t-\tau_1), c_1(t-\tau_1) > + \beta s_2(t-\tau_2) < c_1(t-\tau_1), c_2(t-\tau_2) > \tag{1.25}$$

In each dot product the symbol s_1 and s_2 is constant during a given symbol, so it can be come out of the dot product. Due to the orthogonality of the codes, the second term is much smaller than the first, and we have

$$< r(t), c_1(t-\tau_1) >\approx \alpha s_1(t-\tau_1) \tag{1.26}$$

where we assume the codes are orthonormal. This shows that the receiver can distinguish users in this way by correlating each signal using the appropriate code. But how do we know τ_1 ? Due to the strong delta-like autocorrelation property, we can discover τ_1 by sending a known sequence, or a header for each transmission, and performing autocorrelations of the received signal with different delays until the result is maximized. In other words, we compute $< r(t), r(t-\tau) >$ for a known signal and vary τ to find the delay that maximizes the dot product. Notice that the strong delta-like auto-correlation property also means that if we receive echoes, we can still use the same procedure to recover our signal since the echoes will correlate weakly. Alternatively, our procedure will find the strongest received "copy" (or echo) of the signal, even if the line-of-sight signal is attenuated.

How do we choose c_1 and c_2? One simple choice is pseudo-random sequences of ± 1. Due to the random placement of the $+1$ and -1, the autocorrelation is approximately zero for any delays between the signal and itself. The same is true for two unrelated pseudo-random sequences. The scheme we just described is known ad Code Division Multiple Access (CDMA), which allows multiple users to share the spectrum. With this framework, we recognize that TDMA achieves orthogonality by blanking the signal so that all signals are non-overlapping in time. This approach, though, requires time synchronization and coordination among the users since any mismatch in the timing results in collitions. Another popular approach is Frequency Division Multiple Access (FDMA), whereby we transmit signals at different frequencies. In fact, we can think of discrete sinusoidal tones as codes which have many similar properties to the ones we just discuss.

Sinusoidal codes are advantageous since they are efficient for transmission in a given medium. Sinusoidal signal waveforms have many desirable properties, such as orthogonality, which allows many users to share spectrum by using different *carrier* frequencies. Many antennas are actually resonant circuits that work most efficiently when driven at or near resonance, so using a sinusoidal carrier is also convenient because it allows one to modulate information onto an appropriate carrier frequency. This is why CDMA systems ultimately still use a carrier frequency for wireless transmission. The difference between FDMA and CDMA is that CDMA systems all use the same carrier frequency (channel) whereas FDMA systems use different channels.

Finally, it should be mentioned that many of today's wireless systems also employ OFDMA, or orthogonal frequency division multiple access. WiFi, LTE and even 5G standards use variants of OFDMA and effectively split a given user's channel into more sub-carriers, and modulate these sub-carriers independently. This is done to simplify the equalization of a wideband signal, a topic that we'll cover in Section 1.9.2.

1.7 Wireless Communication Systems and Standards

1.7.1 A Brief History of Wireless

March 10 of 1876 Alexander Graham Bell uttered the words, “Mr. Watson, come here, I want to see you," and Watson heard each word distinctly, the first voice transmission over wires. But even with the ability to transmit voice over wires, the dominant applications of wired communication were telegraphy, essentially digital communication of text using a finite symbol through Morse / Vail Coding. To understand the power of wireless communication, consider the difficulty in wiring different regions of the world separated by oceans. The Transatlantic Cable, competed in 1866 (see Fig. **??**), required thousands of miles of cable to be reeled onto two ships which met in the middle of the Atlantic ocean. In the process of trying to stitch the cables together, one of the cables dropped into the ocean and was nearly lost had it not been recovered by some heroic efforts. This cable operated for a few days and then failed, and understanding how to properly send signals down a long cable led to advancements in the theory of transmission lines.

Early wireless communication started with wireless telegraphy, demonstrated and commercialized by early pioneers such as Marconi, Papov, and Nikola Tesla. Interestingly, the early communications were broadband and digital, utilizing spark-gap generators which emitted broadband pulses. In 1891 Nikola Tesla demonstrated wireless transmission of signals and he suggested wireless telegraphy as an application (Fig. **??**). In November 1894, the Bengali Indian physicist, Jagadish Chandra Bose, demonstrated publicly the use of radio waves in Calcutta, but he did not file for a patent as he was a pure scientist (Fig. **??**). Interestingly, he demonstrated mm-wave transmission even though all early radios were using very long wavelengths. Even though the invention of the wireless radio was a hotly debated topic, Marconi is widely recognized as an early inventor (Fig. **??**), although perhaps he played a more important role in commercializing the radio. In 1895 he sent signals 1.5 km and amazingly, his first transatlantic transmission was achieved in 1902.

In 1912, the RMS Titanic sank in the northern Atlantic Ocean. Wireless radio transmissions (telegraph) were used to report the ship's location. Britain's postmaster-general summed up, referring to the Titanic disaster, “Those who have been saved, have been saved through one man, Mr. Marconi...and his marvelous invention."

Reginald Fessenden invented amplitude-modulated (AM) radio, whereby the envelope of an RF carrier is modulated smoothly in synchrony with a sound wave. On Christmas Eve 1906, Reginald Fessenden made the first radio audio broadcast, from Brant Rock, MA. Ships at sea heard a broadcast that included Fessenden playing O'Holy Night on the violin and reading a passage from the Bible. What is amazing about this story is that the ships were carrying standard receivers used to detect and play telegraph transmissions, not audio. But the same detectors were able to turn the smooth variations in Fessenden's AM modulation waveform into an audible sound (see Fig. **??**a). I'm sure this spooked more than one sea captain.

AM was revolutionary in another way as well. When spark gap generators were employed, virtually the entire usable spectrum was covered by each pulse (generated by a spark), and so interference was a serious issue. But with AM, the amplitude of the signal is modulated in a linear way, so the bandwidth of the audio signal rides on top of the carrier, at most doubling the bandwidth (double-sideband AM). This allowed more than one station can send signals, which led to the popular concept of frequency channels that we use to this day.

The dominant telegraph company of the time was Western Union. They had a monopoly on telegraphy and they dismissed telephony and radio. Telegraphy gave way to audio transmission, mainly phone lines and broadcast radio. Frequency modulation (FM) was invented by Armstrong in 1935 (**??**b). FM has greater noise immunity than AM but requires more bandwidth. Today most high quality audio uses FM (or digital transmission) whereas most AM stations are primarily used for voice and not music.

Digital Wireless

By sampling a signal and quantizing it (turning it into finite precision numbers), we can easily store it using digital technology and we can also transmit it digitally. Audio signals, for example, need to be sampled at about 20,000 times per second and with a resolution of around 18-bits to completely retain the fidelity of the signal (for the "golden" human ear).

Today information is still transmitted with AM and FM, but the amplitude and phase of the signal are mapped into a finite alphabet. These digital signals are more noise immune and can be coded (guarded) to prevent, correct, and detect errors in transmission. Digital signal processing, shown in Fig. **??**, allows us to manipulate digital data in very complex ways, which plays a central role in modern communication. Analog signals are not directly mapped to bits for transmission, but undergo compression and coding, and "predistortion"[4] to minimize the chance for noise and interference to disrupt communication. When a receiver captures bits, it can similarly undo these operations, but also perform equalization to minimize the impact of multi-path propagation (see Sec. 1.9.2), or the fact that wireless signals can travel from the transmitter to the receiver through a rich and diverse set of paths, all arriving at different times, which can potentially cause inter-sybmol interference (ISI) if the variation in the travel time is comparable to the bit period.

In some ways wireless communication has come full circle, going back to digital transmissions, and perhaps even the use of ultra-wideband (UWB) in certain applications. The difference, of course, is that today's digital signals can represent not only text, like telegraph, but also sounds, video, and increasingly data and data about data.

1.8 Spectrum Sharing

Since many users are sharing the same channel, we must contend with interference and come up with a good mechanism to share spectrum. Various techniques have been invented to do this, including frequency channelization, known as Frequend Division Multiple Access (FDMA), time domain multipe access (TDMA), and code division multiple access (CDMA).

[4] The concept of predistortion is a process whereby the data is distorted on purpose to compensate for distortion experienced in the transmitter chain due to unwanted non-linearity in the radio blocks, predominantly the power amplifier.

1.8.1 Choice of Carrier Frequency

Before regulation, anyone could transmit at any desired frequency, which led to a lot of unwanted interference. To resolve this, government agencies (FCC in the U.S.) manage the spectrum, giving licenses to operators in a given region (see Fig. 1.11).

The choice of carrier frequency is both technical and political. More practical technical considerations include the propagation characteristics and the antenna size. For exmaples, frequencies below 1 GHz tend to propagate longer distances and can easily bend around corners (diffraction) whereas signals above 10 GHz take a more direct line-of-sight path, with reflections attenuated more, giving rise to a less diffuse path for radio waves.

Another consideration is bandwidth. As the carrier frequency is moved higher, it's easier to pack more data into a fixed fractional bandwidth. Most information sources are baseband in nature, where we arbitrarily define the bandwidth (BW) as the highest frequency of interest. This usually means that beyond the BW, the integrated energy is negligible compared to the energy in the bandwidth. The bandwidth of some common signals:

- High fidelity audio: 20 kHz (0.72 Mb/s when sampled at 18 bits)
- Uncompressed analog video: 10MHz
- Uncompressed digital video: 1.68 Gb/s (1080×1080 resolution, 3 color channels, 8 bits of color depth, 60 frames per second)
- 802.11 b/g WLAN: 22MHz (peak data rate of 54 Mb/s)

Some common carrier frequencies include the FM band, UHF bands for digital TV and mobile communication, the ubiquitous bands from 1-5 GHz "sweet spot" that balance antenna size with decent radiation and propagation characteristics.

- 100 MHz, FM radio
- 600 MHz, UHF television
- 900 MHz, 1.8 GHz, cellular band
- 2.4 GHz, 5.5 GHz WLAN
- 3-10 GHz, proposed "ultra-wideband"
- 60 GHz unlicensed band

Given the scarcity of free spectrum in the 1-5 GHz range, there's a push to move to higher frequencies, even behond 10 GHz and up to perhaps 100 Ghz, to allow high data rate communication for medium to short ranges. For example, today's fourth generation standard (4G) are bassed on Long Term Evolution (LTE) and operate up to 6 GHz. The fifth generation standard (5G) may use mm-wave bands from 28 GHz - 100 GHz.

1.8.2 FCC Allocation

1.8.3 Cellular (Mobile) Networks)

Let's advance from the early days of AM and FM broadcasting all the way to the year 198X when Motorola engineers demonstrated the first cellular communication system. The modulation scheme was analog FM, and even though walkie-talkies were popular at the time, which allowed peer-to-peer connections in a closed network, a cellular phone system allowed people to call mobile terminals by using the normal phone network. To make this work, basestations were laid out in a grid to cover a geographical. Each basestation has a coverage area, loosely speaking a circle (or a hemisphere) determined by the maximum power transmission and antenna gain. Such a coverage map is often shown with a hexagon coverage area, but in fact the coverage of each base station is more like a circle, which must then coincide to avoid coverage holes. As users move around, they associate with a basestation using a digital control channel. To reach a mobile user, the phone network passes the call to the mobile network, which can then find the user and ring his or her phone. Since coverage is limited, a fixed number of frequency channels can be reused as long as the cells are not adjacent, and so in principle with a fixed frequency allocation of say 10 MHz, if each

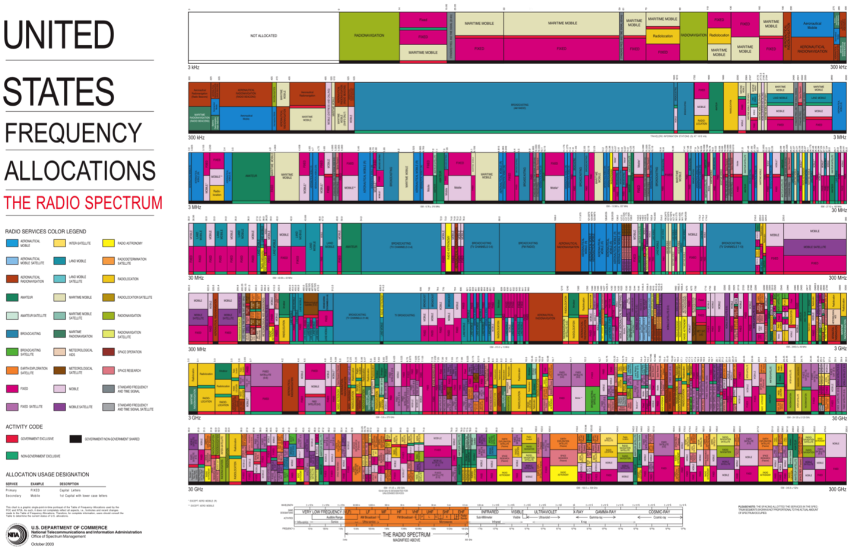


Figure 1.11: How the spectrum is allocated by the FCC in the U.S.

audio channel is say 50 kHz wide, then 200 distinct channels are available, which can be shared among the cells.

Early Analog Systems

The earliest analog mobile phone network to gain an international footing as the AMPS system, or Analog Mobile Phone System (see Fig. 8.9). This system used FM modulation, relatively high power levels (up to 4W), and were primarily voice networks. The AMPS system was a full-duplex system, meaning that the phones could transmit and receive simultaneously. To avoid self-interference, the transmit and receive frequency bands were separated to allow a high isolation passive filter to attenuate the transmit signal from reaching the receiver chain. This filter is called a duplexer.

Digital Modulation

In analog modulation, we vary the amplitude and/or the phase of a sinusoidal carrier signal with our information signal, such as an audio signal. In digital modulation, we also change the amplitude and/or phase of the signal, but use a discrete set of levels to encode the information.

For example, the simplest digital modulation scheme is on-off keying (OOK), a scheme whereby we turn off the transmitter to send a “0” and turn it back on to send a “1”. This is a very simple scheme, but it is also very inefficient use of spectrum. To first order, if we switch our signal at a rate B, then we have B bits per second of datarate and an occupied bandwidht of B (in reality the bandwidth is higher due to the sidelobes of the sinc function, but this is a simple approximation).

For a fixed bandwidth B, we can increase the datarate by using a more complex modulation scheme. Now imagine if we use a more sophisticated scheme, say transmitting multiple amplitude levels to represent the data. Then if we switch at a baud rate B, the datarate is higher by a factor of $\log_2(M)$, where M is the number of levels. For eight levels, we gain 3 bits per Hertz of bandwidth efficiency.

What prevents us from doing this more and more to get more and more bandwidth? There are two limitations, one given by noise, and the other given by the non-linearity of the components in the radio, particularly the transmitter. The limit of noise is easy to understand since if the amplitude levels get too close together, then they are smeared out by noise and we cannot distinguish different symbols. More subtle is the impact of the transmitter power amplifier. Power amplifiers introduce distortion into the signal that also acts like noise and makes it difficult to distinguish transmitted symbols. For these reasons, we cannot pick an arbitrarily complex modulation scheme to increase the datarate.

Second Generation (2G) Systems

The second generation of mobile communication moved to digital modulation and one of the most popular systems, even in use today, is the Global System Mobile (GSM), which uses Minimum Phase Shift Keying (MPSK) modulation to vary the phase of an RF carrier smoothly. Each channel occupies 200 kHz of bandwidth, again with relatively high power levels of about 1W - 2W, and many frequency bands from 800 MHz to 2 GHz were allocated in the US and internationally. In contrast to AMPS, GSM used time-division multiplexing (TDM), meaning that users were allocated a time interval for communication and so the same channels were shared between users. GSM also used time division duplexinig (TDD), so the transmission and reception occurred at separate times, alleviating the need for stringent duplexing. Data communication became more prevalent, although the data speeds were painfully slow.

Third Generation (3G) Systems

With the explosion of wireless communication and the pressing need to support more data communication, the third generation systems strived for more spectral efficiency. For these reasons, 3G systems employed moderately more complex modulation schemes to support data networks, but the systems were still very much built around the telephone network. One big change introduced in 3G was the use of Code Division Multiple Access (CDMA), a technique that encodes multiple independent data streams using the same bandwidth. The concept is exaplained by assuming that each symbol to be transmitted is multiplied by an (approximately) orthogonal set of codes. Say two users want to communicate signals b_1 and b_2, and each uses an independent code $c_1(t)$ and $c_2(t)$, thus both transmit

$$s_1(t) = c_1(t) \cdot b_1(t) \cos(\omega_0 t)$$

$$s_2(t) = c_2(t) \cdot b_2(t) \cos(\omega_0 t)$$

Notice that each user is re-using the same spectrum since both carrriers are centered at ω_0. Now at the receiver, ignoring multipath and filtering, both signals experience the channel gain H

$$r(t) = H \times (s_1(t) + s_2(t))$$

To downconverted these signals to baseband, we once again multiply by the carrier but note that both signals are demodulated to the same baseband

$$v(t) = H \times (s_1(t) + s_2(t)) \cdot \cos(\omega_0 t)$$

$$v(t) = H(c_1 b_1 + c_2 b_2) \cos^2(\omega_0 t)$$

Low-pass filtering this signal, we have

$$v'(t) = H(c_1 b_1 + c_2 b_2)$$

Now we can take advantage of the orthogonality of the codes. If we correlate the output with c_1 over a period of a symbol, we have

$$c(t) =< c_1(t), v'(t) >= H < c_1, c_1 > b_1 + H < c_1, c_2 > b_2 \approx H < c_1, c_1 > b_1$$

In the above calculation we assume that b_i is constant over the period of the correlation and then we use the fact that $< c_1, c_2 > \approx 0$. In this way, we recover user 1's transmission. In a similar way, we can also recover user 2's transmisison, and this concept can be extended to multiple users.

One challenge with CDMA is that the codes are never going to be perfectly orthogonal, which means that the cross-correlation term in the above calcalution is not zero but some small error term which acts like noise. If we have more users, more cross-correlation terms will increase this noise floor and limit the efficacy of the communication. Notice that this "noise" increases in proportional to the signal power, and for this reason, in CDMA systems there is a need to implement strict power control to limit the transmitter output power to the minimum required to acheive communication. In older 2G systems, the transmitter were allowed to "blast" away to get the maximal coverage and very little if any power control was used. But in CDMA it became essential to limit the power.

Fourth Generation (4G) Systems

The fourth generation of mobile commucation, in wide use today, is the Long Term Evolution (LTE) system, which is much more of a data-centric network. Wider band channels, and much more sophisticated modulation schemes are used to to increase the throughput and to communication effectively in the presence of multi-path propagation and frequency selective fading. LTE systems occuty bands from 700 MHz all the way up to 2.6 GHz and beyond, with different bands used in different parts of the world.

FDD/TDD

Carrier aggregation

SC-OFDM

Fifth Generation (5G) Systems

massive mimo

1.8.4 Local Area Networks

802.11a/b/g (WiFi)

OFDM

802.11n

MIMO

802.11ad

60 GHz

1.8.5 Personal Area Networks

Bluetooth

UWB

1.8.6 Locationing and Positioning

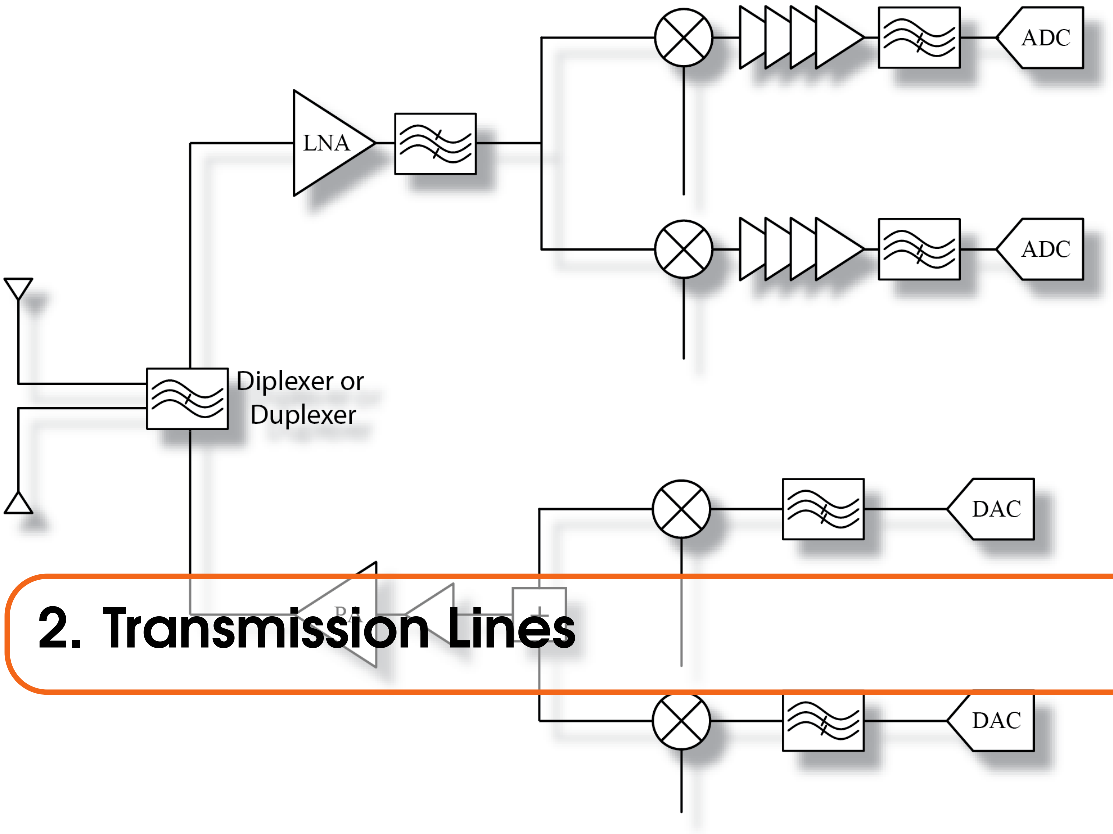
GPS

1.9 Wired versus Wireless Propagation

1.9.1 Line of Sight Propagation

1.9.2 Multi-Path Propagation

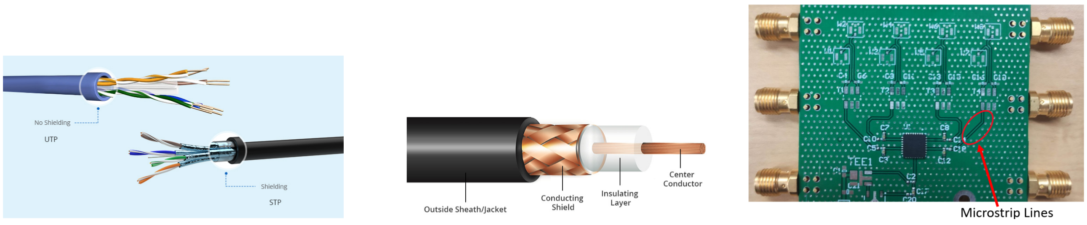
1.10 Narrowband versus Ultra Wideband



2. Transmission Lines

Transmission lines are essentially long cables or long traces of wires, often carrying high frequency signals. Common examples include the cables connecting your laptop to a monitor, or the phone wires (“twisted pair") coming into your house from the central office or Ethernet cables (see Fig. 2.1a), or the long coaxial cables that transmit data/video into homes using Cable Modems (see Fig. 2.1b). Other less recognizable examples include traces on a PCB (see Fig. 2.1c) or even very “long" traces inside a chip. What constitutes “long" is not so much the length of the cable, but the “electrical length", which we'll explore in the following chapters. Electrical length is roughly the length normalized by the wavelength of the highest frequency of interest. Power lines are also referred to as transmission lines, but at 60 Hz they are usually not “electrically" long and can be modeled as lumped wires. That's because the wavelength at 60 GHz is very long ($\lambda = c/f$), since the speed of light is so fast.

Transmission line behavior represents a true departure from lumped circuit theory. This is most poignant when we consider the input impedance of a quarter-wavelength shorted transmission line. Circuit theory cannot account for the fact that the input impedance is actually infinite, or an open, rather than a short. But circuit theory can be used as a foundation to understand these effects. This



(a) (b) (c)

Figure 2.1: (a) Twisted-pairs are commonly used in telephony and data cables such as Ethernet. (b) Coaxial cables, or “coax" cables provide more isolation and bandwidth. (c) PCB traces are transmission lines, often using the ground plane as the return path.

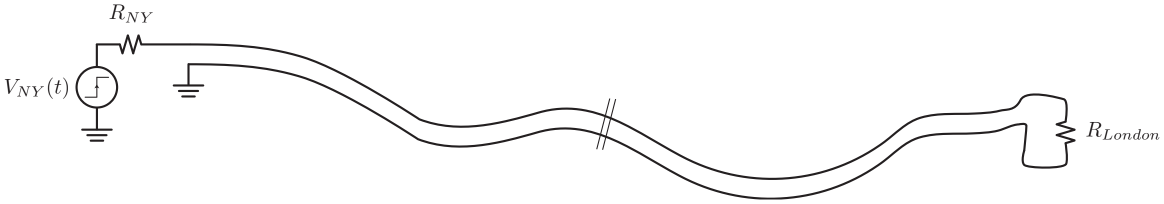


Figure 2.2: A long transatlantic cable from New York to London.

comes about when we expand our circuit theory to account for the *distributed* nature of the circuit elements.

In circuit theory we implicitly assume that all signals travel throughout the circuit infinitely fast. In reality, signals cannot travel faster than the speed of light. Thus there is always a delay from one point in a circuit to another point. In fact circuit theory is strictly valid in the limit of a truly *lumped* circuit, or a circuit with zero physical dimension.

Thus distributed effects become important when circuits become electrically large. In time-harmonic problems, we can relate the speed of light to the wavelength, hence the distributed effects become important when physical circuit dimensions approach the wavelength of electromagnetic propagation in the medium λ. For example, a circuit with dimensions of 3 cm in free space is electrically large when the operating frequency approaches $f = c/(.1\lambda)$ or about 10 GHz. We have arbitrarily used $\lambda/10$ to denote this boundary whereas in reality the cutoff is application dependent. If the waveforms are non-sinusoidal, then the high frequency spectral content of the waveform is important. For instance, in digital applications the clocking frequency may be low, but due to the fast edges of the waveform (risetime and falltime), there is appreciable energy at higher harmonics.

2.1 Distributed Properties of a Cable

Consider a long cable, say a transatlantic cable running from London to New York (5585 km) shown in Fig. 2.2. The cable has a uniform cross-section, and if we ignore any non-uniformity caused by the bends, we can consider it a uniform two-line wire. How do we model such a cable? At very low frequencies, we can think of this wire as a series inductor (with loss). As shown in Fig. 2.3a, we can calculate the equivalent series impedance by shorting the end of the wire and measuring the input impedance. But also, as shown in Fig. 2.3b, if we leave the other end of the cable open, certainly there is a substantial shunt capacitance associated with the cable. If the cables are unshielded and uninsulated, then due to the conductivity of sea-water, there is also a shunt conductance associated with the cable, even when left open.

Since the cable is very long and the delay associated with signal propagation is significant ($t_d \approx 20\,\mu$s in air over the same distance), we know that a lumped circuit model will not properly account for the behavior of the line. The large series inductance would render the wire an open circuit, even at very low frequencies, whereas the large shunt capacitance would makes the wire a short circuit. So how do signals propagate through this cable? Before tossing lumped circuit theory in the can note that any small section of the wire still behaves like a lumped circuit. For instance, if we're concerned with signal propagation up to say 1 GHz, we can break up our long cable into short sections $\ell_{sec} \ll \lambda = 30$ cm, and use circuit theory on the resulting large number of cascaded lumped circuits.

We can now take the total series impedance and shunt admittance and break it up into $N = \ell/\ell_{sec}$ section, as shown in Fig. 2.4. For simplicity, we have ignored loss in the circuit. For instance with $\ell = 1$ cm, we would need about 1 billion inductors and capacitors. Even with today's powerful computers, running a SPICE simulation with that many elements would tax our computers. So how

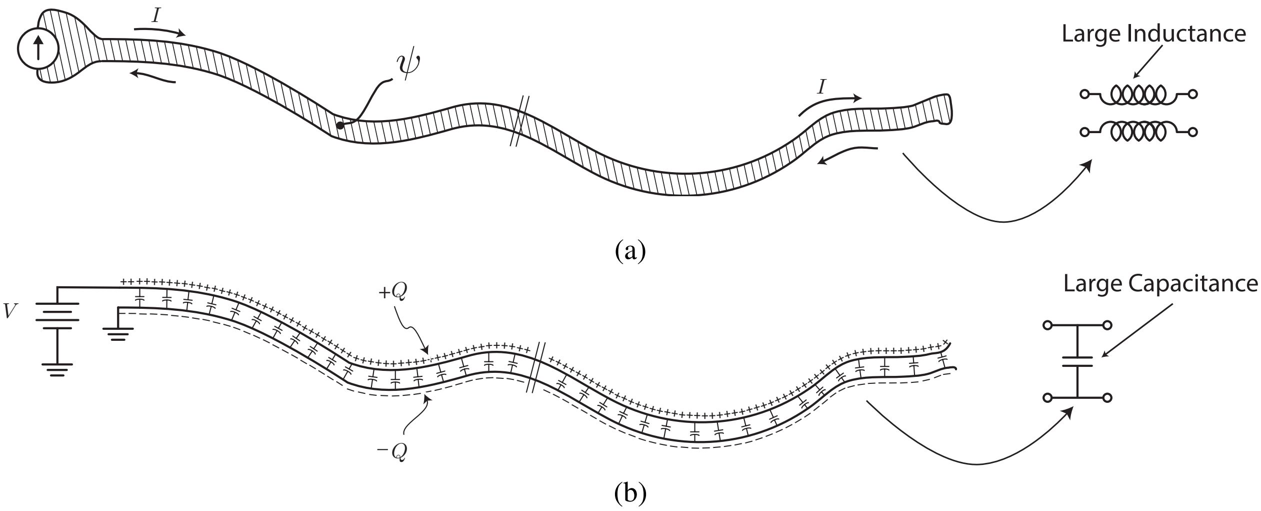


Figure 2.3: (a) The series impedance (inductance and resistance) associated with the transatlantic cable can be measured at very low frequencies by measuring the short-circuit input impedance of the line. (b) Likewise, the shunt admittance (capacitance and conductance) can be measured using the open-circuit admittance at very low frequency.

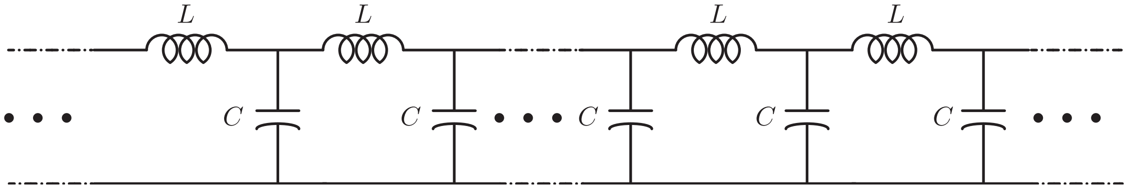


Figure 2.4: An LC ladder network representation of the transatlantic cable (neglecting cable loss). For a given frequency the required number of sections can be computed to obtain accurate results.

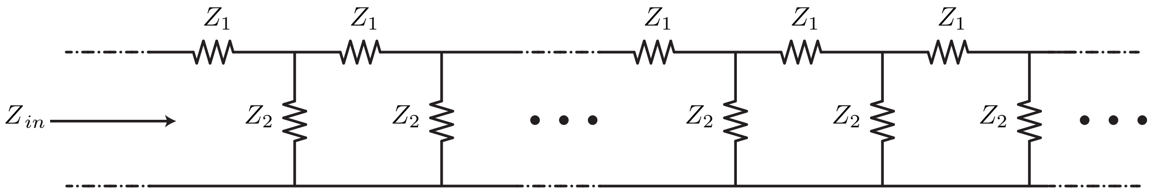


Figure 2.5: An infinite ladder network with regular structure of series impedance Z_1 shunt impedance Z_2.

did people solve this problem when the first Transatlantic Cable was laid out in 1857?

2.1.1 An Infinite Ladder Network

We begin our study of transmission lines by first studying an infinite lumped ladder network shown in Fig. 4.14. It is interesting that we can find the input impedance of such a network (often a freshman physics problem). Simply observe that since the latter network is infinite, addition of a single section to the *front* should not alter the impedance. With this observation one can shown that

$$Z_{in} = Z_1 + Z_2||Z_{in} \tag{2.1}$$

or

$$Z_{in}^2 - Z_1 Z_{in} - Z_2 Z_2 = 0 \tag{2.2}$$

As shown in Fig. 2.4, suppose now that $Z_1 = j\omega L$ and $Y_2 = j\omega C$. Then the input impedance of such a line is

$$Z_{in} = \frac{j\omega L}{2} \pm \sqrt{-\frac{(\omega L)^2}{4} + \frac{L}{C}} \tag{2.3}$$

We would now like to make the leap from lumped to distributed. As such, we will assume that each inductor in the ladder is very very small, in fact, infinitesimal in size. Therefore, for any finite frequency, the input impedance degenerates to

$$Z_{in} \approx \sqrt{\frac{L}{C}} \tag{2.4}$$

This is a very important and profound result. The input impedance is positive and real! Notice that we started out with strictly reactive elements and managed to construct a circuit with positive real input impedance. This final result is puzzling because the power dissipated by such a network is proportional to $\Re(Z_{in})$. But since each section of the ladder is incapable of dissipating power, where does the energy go?

We will answer this question in due course but for now we will hint at the solution. The energy is absorbed by the network because it is infinite in extent. The energy pumped into the network keeps vacillating from inductive to capacitive energy as it travels through the ladder network. Since there are an infinite number of capacitors and inductors to charge and discharge, the process goes on indefinitely.

Observe that if we choose to terminate the ladder network at any point with an impedance of Z_{in}, then the above results remain valid. The behavior of a finite ladder section terminated with Z_{in} is indistinguishable from that of the infinite network. In the case of a distributed line, the termination resistance is real and given by $Z_0 = \sqrt{L/C}$. The energy injected into the network, therefore, is absorbed by this resistor.

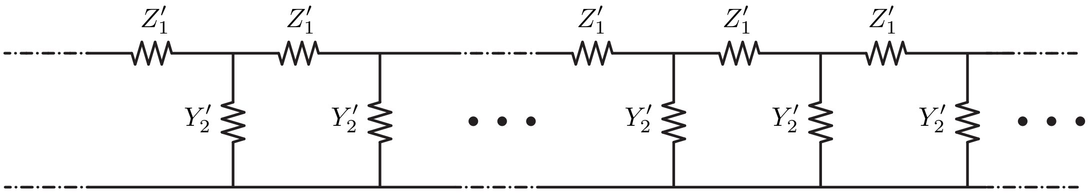


Figure 2.6: A distributed ladder network with a series impedance per unit length Z_1' and a shunt admittance per unit length of Y_2.

2.1.2 Transmission Lines as Distributed Ladder Networks

We can analyze the two-wire transmission line shown in Fig. 2.2 using the concept of distributed circuits. We note that this two-wire line stores both magnetic and electrical energy everywhere along the line in a distributed fashion. In other words, we cannot say that in some region the magnetic energy storage dominates and thus the line behaves inductively whereas in another region the electrical energy storage dominates and thus the lines behaves capacitively. Thus we now use a purely distributive circuit perspective shown in Fig. 2.6, a distributed ladder network with series impedance Z_1' and shunt admittance Y_2'. The prime denotes that the element in question is distributed, in other words there is Z_1' impedance per unit length.

To approximately find Z_1', we ignore the effects of the shunt admittance by focusing on the magnetic energy stored in the line. To do this, short the line at a distance ℓ from the input, as shown in Fig. 2.3a, and the line behaves like an inductor with impedance Z_1 accounting for its magnetic energy storage and loss. Since the line is uniform, we may define the impedance per unit length

$$Z_1' = \frac{Z_1}{\ell} \tag{2.5}$$

Similarly, if we break the two-wire line at a distance ℓ and keep the end open, as shown in Fig. 2.3b, and only consider the electrical energy storage in the line, the line behaves capacitively. And so we define the admittance per unit length

$$Y_2' = \frac{Y_2}{\ell} \tag{2.6}$$

In general, the electrical and magnetic energy storage are intermingled and thus we must distribute the series and shunt impedances uniformly along the line.

So far we have used the two-wire transmission line as an example. But our discussion applies to any structure with a uniform two-dimensional cross section. Common examples are shown in Fig. 2.7. For now we restrict our discussion to structures with two conductors, where one conductor is chosen as the reference plane. If more than one conductor is present, it is assumed that all but one conductor is at the same reference or ground potential. This is true, for instance, in the coplanar line shown in Fig. 2.7d, where the conducting planes on the left and right of the conductor are the ground return path for the signal conductor situated in the middle.

If there's only one conductor, then we cannot form a loop for current flow unless we define the return path. For example, earch “ground" can act as a return path if signals are reference to earth ground. This is inadvisable for many reasons, and it's much more common to employ two conductors to realize a transmission line. Even if more than one signal is being transmitted, such as a calbe bundle, it's advisable to separate the ground connections rather than sharing the ground. This reduces crosstalk, or undesired coupling from one transmission line to the others.

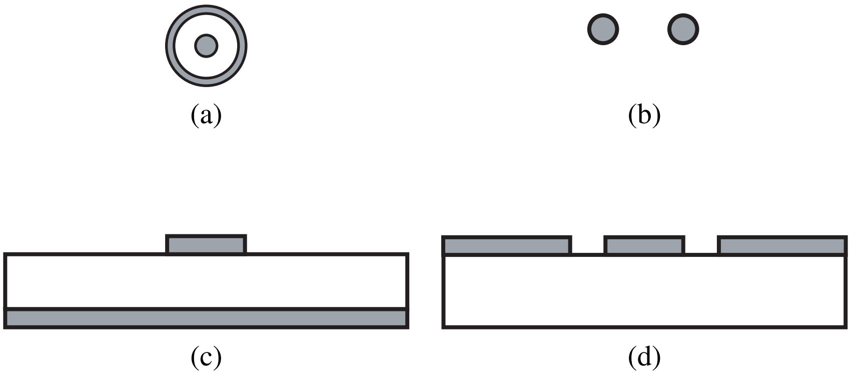


Figure 2.7: Several examples of common transmission line structures. (a) A familiar coaxial transmission line is ubiquitous in video applications. (b) A pair of wires are employed in some antenna feedlines and more importantly in *twisted* form as telephone and data cables. (c) A microstrip line and (d) a coplanar line are common structure for building transmission lines over a printed-circuit board (PCB) or in a Microwave Integrated Circuit. In the coplanar structure the middle conductor carries the signal whereas the outer conductors form the ground return path.

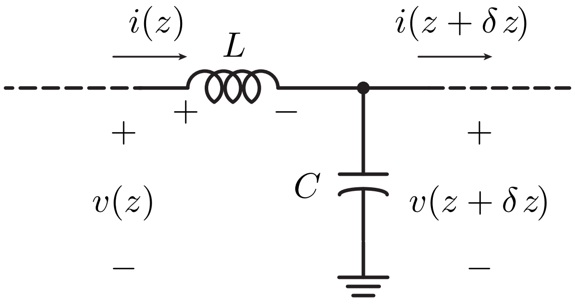


Figure 2.8: Applying KCL and KVL to a short section of the line.

2.2 Telegrapher's Equations

Given that the first application of transmission lines were to convey telegraph signals over a long distance, the most important equations bear this name[1]. Going back to the LC line shown in Fig. 2.8, we can use circuit theory to analyze a very small section δz. As we take the limit $\delta z \to 0$, the model becomes exact. We defined inductance and capacitance per unit length $L' = L/\ell$, $C' = C/\ell$. For now we will ignore loss and only focus on L' and C'.

2.2.1 KCL and KVL for a Small Section

From KCL we see the current flowing into a section has two parts, one part that flows to the second conductor (ground) via C' and a part that passes to the next section through L':

$$i(z) = \delta z C' \frac{\partial v(z)}{\partial t} + i(z + \delta z)$$

[1] Telegrapher was not a famous scientist or engineer !

Likewise, from KVL we see that the voltage difference between two sections is given by the voltage drop across the inductor L':

$$v(z) - v(z+\delta z) = \delta z L' \frac{\partial i(z+\delta z)}{\partial t}$$

Take limit as $\delta z \to 0$ we arrive at "Telegrapher's Equations".

$$\lim_{\delta z \to 0} \frac{i(z) - i(z+\delta z)}{\delta z} = -\frac{\partial i}{\partial z} = C' \frac{\partial v}{\partial t} \tag{2.7}$$

$$\lim_{\delta z \to 0} \frac{v(z) - v(z+\delta z)}{\delta z} = -\frac{\partial v}{\partial z} = L' \frac{\partial i}{\partial t} \tag{2.8}$$

2.3 Derivation of Wave Equations

We have two coupled equations and two unkowns (i and v). Of course the voltage and current are couopled. We can reduce it to two de-coupled equations if we first note that:

$$\frac{\partial^2 i}{\partial t \partial z} = -C' \frac{\partial^2 v}{\partial t^2} \tag{2.9}$$

$$\frac{\partial^2 v}{\partial z^2} = -L' \frac{\partial^2 i}{\partial z \partial t} \tag{2.10}$$

Since the order of partials can be changed for functions satisfying some continuity conditions (see Clairaut's theorem), we have:

$$\frac{\partial^2 v}{\partial z^2} = L'C' \frac{\partial^2 v}{\partial t^2} \tag{2.11}$$

The same equation can be derived for current:

$$\frac{\partial^2 i}{\partial z^2} = L'C' \frac{\partial^2 i}{\partial t^2} \tag{2.12}$$

We see that the currents and voltages on the transmission line satisfy the one-dimensional Wave Equation. This is a partial differential equation that appears in many different contexts, from sound waves to water waves to vibrating strings and to electromagnetic waves (light). To find the solution, we must satisfy not only the equation but as the boundary conditions and the initial condition.

2.3.1 Wave Equation Solution

Consider the candidate function $f(z,t) = f(z \pm vt) = f(u)$. We can take partial derivatives using with respect to z and t using u. Let's start with spatial derivatives:

$$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial u} \frac{\partial u}{\partial z} = \frac{\partial f}{\partial u} \tag{2.13}$$

So taking derivatives with respect to z is the same as u:

$$\frac{\partial^2 f}{\partial z^2} = \frac{\partial^2 f}{\partial u^2} \tag{2.14}$$

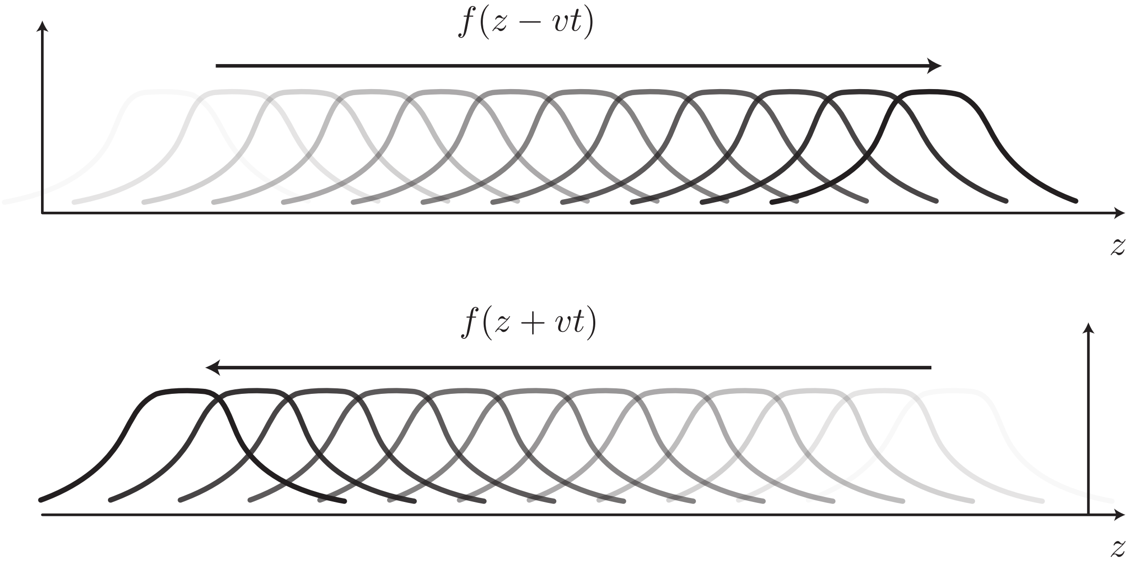


Figure 2.9: The functions $f(z \pm vt)$ represent a propagating waveform to the right $(-)$ or left $(+)$.

Next consider time derivatives:

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial t} = \pm v\frac{\partial f}{\partial u} \tag{2.15}$$

We see that we can take derivatives with respect to time by taking derivatives with respect to u and multiplying by $\pm v$:

$$\frac{\partial^2 f}{\partial t^2} = \pm v\frac{\partial}{\partial u}\left(\frac{\partial f}{\partial t}\right) = v^2\frac{\partial^2 f}{\partial u^2} \tag{2.16}$$

Putting this all together:

$$\frac{\partial^2 f}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2} \tag{2.17}$$

We find that this candidate function satisfies the wave equation.

Wave Motion

Why is it called the “Wave Equation"? Well, we just found that any function that “moves" or propagates satisfies the wave equation, as shown in Fig. 2.9. The general voltage solution is therefore given by

$$v(z,t) = f^+(z - vt) + f^-(z + vt)$$

where $v = \sqrt{\frac{1}{L'C'}}$. The parameter v is very important and considered next.

Wave Speed

Speed of motion can be deduced if we observe the speed of a point on the waveform. If we hold the argument of $f(u)$ constant, and follow that point, we can derive the speed:

$$z \pm vt = \text{constant} \tag{2.18}$$

To follow this point as time elapses, we must move the z coordinate in step with time. This point moves with velocity

$$\frac{dz}{dt} \pm v = 0 \tag{2.19}$$

This is the speed at which we move with speed $\frac{dz}{dt} = \pm v$, and so v is the velocity of wave propagation to the right or left.

2.3.2 Current / Voltage Relationship

Since the current also satisfies the wave equation, it also has the solution that takes the following general form:

$$i(z,t) = g^{+}(z - vt) + g^{-}(z + vt) \tag{2.20}$$

Recall that on a transmission line, current and voltage are related by

$$\frac{\partial i}{\partial z} = -C' \frac{\partial v}{\partial t} \tag{2.21}$$

For the general function this gives

$$\frac{\partial g^{+}}{\partial u} + \frac{\partial g^{-}}{\partial u} = -C' \left(-v \frac{\partial f^{+}}{\partial u} + v \frac{\partial f^{-}}{\partial u} \right) \tag{2.22}$$

Since the forward waves are independent of the reverse waves, we can equate each term individually:

$$\frac{\partial g^{+}}{\partial u} = C' v \frac{\partial f^{+}}{\partial u} \tag{2.23}$$

$$\frac{\partial g^{-}}{\partial u} = -C' v \frac{\partial f^{-}}{\partial u} \tag{2.24}$$

Within a constant we have related the voltage and current waves:

$$g^{+} = \frac{f^{+}}{Z_0} \tag{2.25}$$

and

$$g^{-} = -\frac{f^{-}}{Z_0} \tag{2.26}$$

where $Z_0 = \sqrt{\frac{L'}{C'}}$ is the "Characteristic Impedance" of the line. The Z_0 of the line will play a very important role when we study the boundary conditions of the line.

A Side Note on Current

Notice that the currents in the forward wave has the same sign

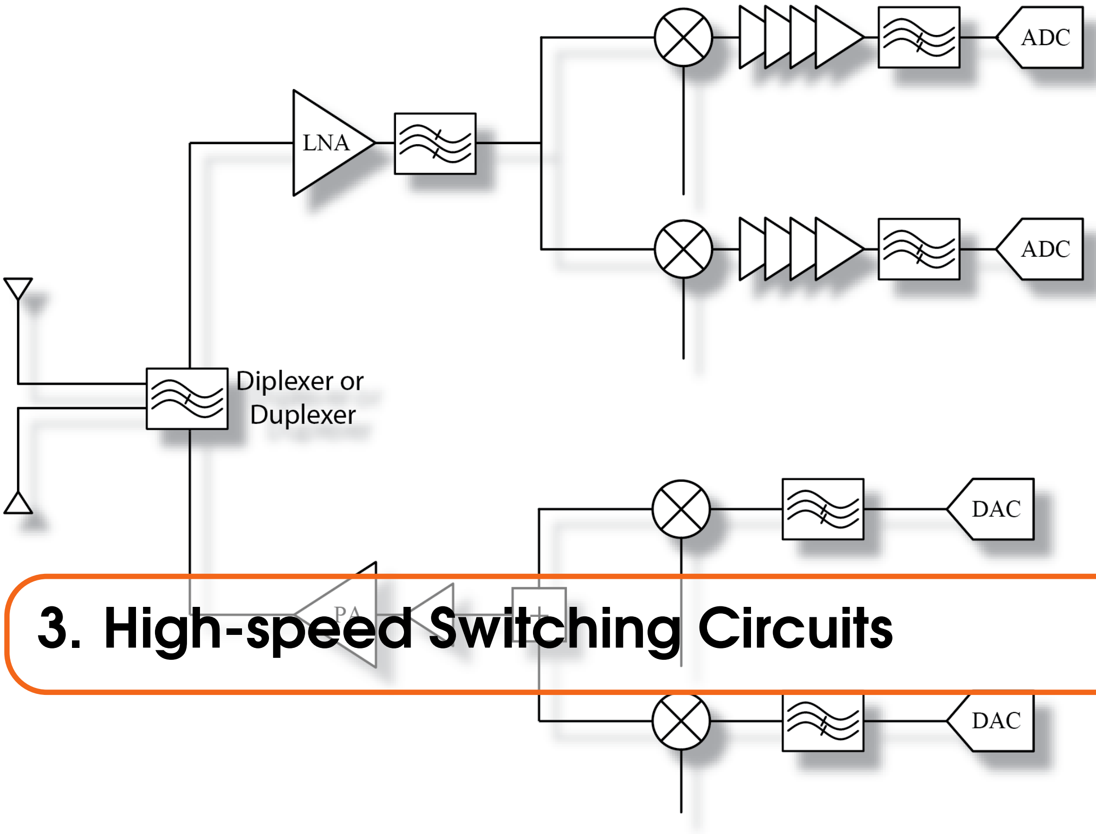
$$g^{+} = \frac{v^{+}}{Z_0} \tag{2.27}$$

But the reverse wave has a negative sign

$$g^{-} = -\frac{v^{-}}{Z_0} \tag{2.28}$$

This is related to the definition of current. If positive charges are moving left, then the corresponding current is negative.

It's important to understand the definition of currents on a transmission line with respect to the two conductors. g^{+} is not the current on the top conductor and g^{-} is not the bottom conductor, it's the current flowing on both conductors, and the magnitude is equal and opposite. So g^{+} and g^{-} flow on both conductors. You may be wondering why the currents are equal and opposite. That's because we only consider two conductors, so the return current has to in the bottom conductor. This is also known as an “odd" mode current, since the top and bottom conductors carry equal and opposite currents. There's also an “even" mode current that we are neglecting for now, and we'll return to this topic later. The even-mode current requires the presence of a third conductor.



3. High-speed Switching Circuits

A lot of this book will be concerned with time harmonic excitation and behavior of passive and distributed elements, in other words we deal explicitly with the frequency domain. This is both convenient as the differential equations simplified to algebra but it is also a common situation. Most RF circuits, for example, are narrowband, and therefore for all practical purposes the input signal is a sinusoid at the carrier frequency. But there are a host of situations where the signals of interest are wideband. This includes digital switching circuits, high speed links, and ultra-wideband circuits and systems. Furthermore, digital circuits are switching increasingly faster, now into the multi-GHz regime. There are also many high-speed links, such as chip-to-chip communication on a PCB. These links are also referred to as SerDes links for serializer/deserializer links, whereby a large number of parallel links are aggregated and serialized into a single high speed link. Since multiple lines are aggregated, the data rates can be quite large, for example as high as 100-200 Gb/s.

These high speed interface circuits drive pulses or spectrally rich waveforms on long board traces. We will treat these different cases in a uniform manner by generically calling these circuit applications “high-speed” switching circuits.

3.1 Transmission Lines

As we discussed in the previous chapter, transmission lines are essentially a pair of cables with a fixed spacing, such as a single trace and a ground plane (microstrip line), or physical wires either held together with a dielectric, or twisted together. Other ubiquitous transmission lines include coaxial cables, with an “inner" conductor (wire) and an outer shield (hollow cylinder), separated by air or a dielectric. Several examples are shown in Fig. 3.1. Why do we treat these wires differently than simple lumped wires ? The reason is that these wires have significant inductance, capacitance, and also loss (covered in the next chapter). The inductance and capacitance is not lumped but distributed along the line.

3.2 Transmission Lines and High-Speed Switching Circuits

The focus of the chapter is switching waveforms on transmission lines. You may wonder why we would use transmission line analysis for switching circuits, especially small on-chip or on-board circuits. Let's take a practical example to get a feel for the problem. Let's say a digital chip has a dimension of less than 1 cm. For a time-harmonic circuit, we know that we may treat most structures as lumped as long as their dimension does not exceed $(1/10)\lambda$. For a SiO_2 transmission line structure (speed of light is about half of the vacuum speed), we equate the chip dimension to one tenth the wavelength $\lambda = 10 \times 1\,\mathrm{cm} = 10\,\mathrm{cm}$, in order to solve for the maximum operating frequency

$$f = c/\lambda = \frac{3}{2} \times 10^8/10\,\mathrm{cm} = 1.5\,\mathrm{GHz} \tag{3.1}$$

For signals with center frequency below about 1 GHz, therefore, we may use lumped analysis.[1]

For a digital waveform, or a switching waveform, we cannot simply use the switching or clock frequency F_{clk} to determine the boundary between lumped and distributed analysis. Recall that a deterministic square wave has a discrete spectrum that decays like $1/N$ with the N'th harmonic frequency. For a randomly switching waveform, we must evaluate the Fourier transform of the autocorrelation function [**proakis**], which results in the well known sinc characteristic. This spectrum is rich in harmonics and therefore a good fraction of the energy will appear at frequencies above the switching rate. A more realistic waveform would have a non-zero rise and fall time. For such a waveform, it can be shown that most of the energy is contained in the spectrum from DC to the knee-frequency $F_{\mathrm{knee}} = 0.5/T_r$, where T_r is the waveform risetime (falltime) [**magic**].

From this we conclude that a digital waveform with knee frequency significantly below 1.5 GHz, can also be treated like a lumped circuit. This corresponds to a risetime of

$$T_r = 0.5/F_{\mathrm{knee}} = 0.5/1.5\,\mathrm{ns} = \frac{1}{3}\,\mathrm{ns} \tag{3.2}$$

If we take the risetime to be 10% of the clock period, then the minimum clock period is $T = 10 \times T_r \approx 3\,\mathrm{ns}$, or $F_{\mathrm{clk}} = 300\,\mathrm{MHz}$. Digital circuits are already exceeding this frequency by an order of magnitude. But since most on-chip circuits are a small fraction of the chip dimension, with the exception of the clock distribution network or a signal that must propagate over a long distance, most on-chip structures can be treated in a lumped manner. But now we see that off-chip structures, the package, board, connectors, and especially cables, are long enough to require full distributed analysis.

Advantages of Transmission Lines

There are many situations as a circuit designer where we must decide between using transmission lines and simple point to point connections [**magic**]. In Fig. 3.1 we show a typical example where two integrated circuits are communicating on a PCB. Let's say that the two chips are close enough where simple wires can be used to connect the ICs rather than transmission lines. Is there any reason to use transmission lines at all in this case?

If we use a "wire" to connect the two chips, then we would keep the length as short as possible to minimize the inductance. If the input impedance of the integrated circuit is capacitive, a common situation, the wire can be modeled by an LC circuit, where the capacitance is due to the wire capacitance summed with the output and input capacitance of the circuit. From chapter 3.8, we know that the step response of this circuit will exhibit ringing if the $Q > 1/2$. Ringing is undesirable

[1] We should be careful to not excite the circuit in *slow wave* modes that have significantly reduced velocity. This occurs, for instance, when waves flow between a conductor with a return current flowing in the Si substrate.

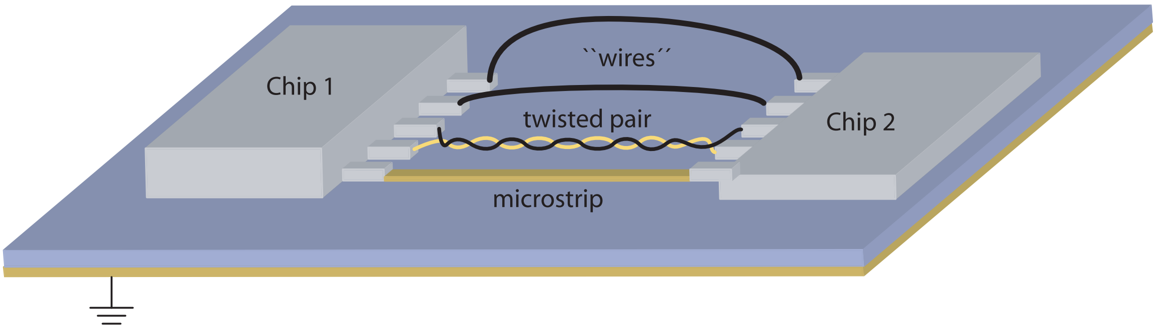


Figure 3.1: Two chips can be connected with wires using the backplane ground reference, or more explicitly with a twisted pair connection or a microstrip transmission line.

because it forces us to slow the clock down to allow the waveforms to settle to the correct values. Alternatively, we can introduce loss into the circuit to lower the Q to eliminate the ringing.

But there are other problems with this circuit as shown. The high inductance wire tends to generate large magnetic fields that will generate a lot of EMI (electromagnetic interference). This means that the circuit may fail to qualify for a production environment. More importantly, the circuit has a large tendency to interfere due to mutual inductance and capacitnace. Any nearby wire experiences cross-talk due to the mutual inductance and capacitance. Charging and discharging currents MdI/dt couple the flux leading to potential problems. Likewise, changing voltages coupled through CdV/dt displacement currents. The crude solution, as always, is to slow down the clock. This of course also means increasing the risetime.

Alternative interconnections are shown in Fig. 3.1. We reduce the wire inductance by placing it closer to the ground plane, thus naturally creating a transmission line structure. The reduction in inductance occurs since ground return current is closer to the signal. If the ground plane is far away, we can use a ground wire in a twisted pair fashion, forming a pseudo-transmission line. While two parallel wires would form a real transmission line, the twisted pair has better isolation. If a ground plane is available, we can use a PCB trace to form a microstrip transmission line. This naturally cuts down the area of the loop and minimizes the magnetic coupling and EMI. But what about the ringing?

In the next sections we shall show that the circuit with the transmission line will also ring unless we properly *terminate* the transmission lines. These transmission line terminations require us to drive resistive loads, which increases the power dissipation of the circuits. But in return our circuits are robust and can be driven to very high clock frequencies.

3.3 Transients on Transmission Lines

Time Domain Voltage/Current Waveforms

In Chapter 2 we found that the currents and voltages on the transmission line satisfy the one-dimensional wave equation. This is a partial differential equation. The solution depends on boundary and the initial conditions. For the lossless case, we found

$$\frac{\partial^2 f}{\partial z^2} = L'C' \frac{\partial^2 v}{\partial t^2} \tag{3.3}$$

We showed that the function $f(z,t) = f(z \pm vt) = f(u)$ satisfies the wave equation. In theory, any waveform can be excited onto a lossless transmission line and it will travel along the line without dispersion.

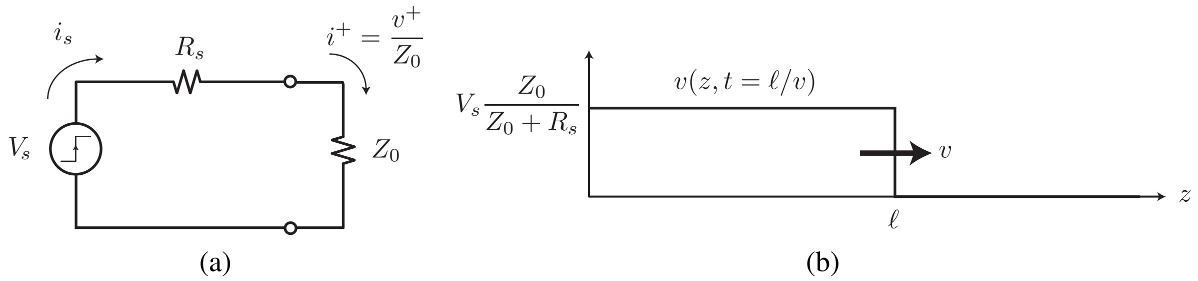


Figure 3.2: (a) Equivalent circuit for a voltage source driving an infinite transmission line. (b) The waveform on the transmission line at time $t = \ell/v$ due to a step function a the source.

The most general voltage solution is therefore two waves, $v(z,t) = f^+(z - \nu t) + f^-(z + \nu t)$, where $\nu = \sqrt{\frac{1}{L'C'}}$. This is the speed at which we move, $\frac{dz}{dt} = \pm\nu$, and ν is the velocity of wave propagation. The current also satisfies the wave equation $i(z,t) = g^+(z - \nu t) + g^-(z + \nu t)$. Recall that on a transmission line, current and voltage are related by

$$g^+ = \frac{f^+}{Z_0} \tag{3.4}$$

and

$$g^- = -\frac{f^-}{Z_0} \tag{3.5}$$

where $Z_0 = \sqrt{\frac{L'}{C'}}$ is the "Characteristic Impedance" of the line.

3.4 Step Function Excitation of an Infinite Line

Consider exciting a step function onto a transmission line. Note that a finite width pulse can be constructed as a sum of two step functions, $p_\tau(t) = u(t) - u(t - \tau)$. The line is assumed uncharged: $Q(z,0) = 0$, $\psi(z,0) = 0$ or equivalently $v(z,0) = 0$ and $i(z,0) = 0$. By physical intuition, we would only expect a forward traveling wave since the line is infinite in extent. The general form of current and voltage on the line is given by

$$v(z,t) = v^+(z - vt) \tag{3.6}$$

$$i(z,t) = i^+(z - vt) = \frac{v^+(z - vt)}{Z_0} \tag{3.7}$$

The T-line looks like a resistor of Z_0 ohms! We may therefore model the line with the equivalent circuit shown in Fig. 3.2a.

Since $i_s = i^+$, the excited voltage wave has an amplitude of

$$v^+ = \frac{Z_0}{Z_0 + R_s} V_s \tag{3.8}$$

It's surprising that the voltage magnitude on the line is not equal to the source voltage. As shown in Fig. 3.2b, the voltage on the line is a delayed version of the source voltage.

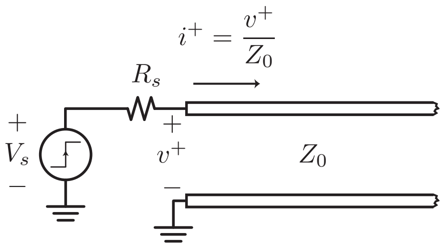


Figure 3.3: The power delivered to the infinite transmission line by a step voltage source.

Energy on Transmission Line

Let's calculate the energy to "charge" a transmission line. The setup is shown in Fig. 3.3. The power flow into the line is given by

$$P^+_{line} = i^+(0,t)v^+(0,t) = \frac{(v^+(0,t))^2}{Z_0} \tag{3.9}$$

Or in terms of the source voltage

$$P^+_{line} = \left(\frac{Z_0}{Z_0 + R_s}\right)^2 \frac{V_s^2}{Z_0} = \frac{Z_0}{(Z_0 + R_s)^2} V_s^2 \tag{3.10}$$

But where is the power going? The line is lossless! The energy stored by a capacitor and inductor is $\frac{1}{2}CV^2/\frac{1}{2}LI^2$. At time t_d, a length of $\ell = vt_d$ has been "charged"

$$\frac{1}{2}CV^2 = \frac{1}{2}\ell C' \left(\frac{Z_0}{Z_0 + R_s}\right)^2 V_s^2 \tag{3.11}$$

$$\frac{1}{2}LI^2 = \frac{1}{2}\ell L' \left(\frac{V_s}{Z_0 + R_s}\right)^2 \tag{3.12}$$

The total energy is thus

$$\frac{1}{2}LI^2 + \frac{1}{2}CV^2 = \frac{1}{2}\frac{\ell V_s^2}{(Z_0 + R_s)^2}\left(L' + C'Z_0^2\right) \tag{3.13}$$

Recall that $Z_0 = \sqrt{L'/C'}$. The total energy stored on the line at time $t_d = \ell/v$

$$E_{line}(\ell/v) = \ell L' \frac{V_s^2}{(Z_0 + R_s)^2} \tag{3.14}$$

And the energy delivered onto the line in time t_d:

$$P_{line} \times \frac{\ell}{v} = \frac{\frac{l}{v} Z_0 V_s^2}{(Z_0 + R_s)^2} = \ell\sqrt{\frac{L'}{C'}}\sqrt{L'C'}\frac{V_s^2}{(Z_0 + R_s)^2} \tag{3.15}$$

As expected from conservation of energy, the results match.

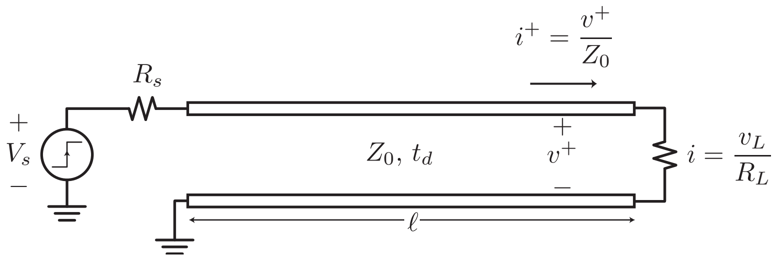


Figure 3.4: A finite section of transmission line terminated with a load resistance R_L.

3.5 Terminated Transmission Line

Consider a finite transmission line with a termination resistance. This setup, shown in Fig. 3.4, is perhaps the most common practical situation. At the load we know that Ohm's Law applies, $I_L = V_L/R_L$. So at time $t = \ell/v$, our step reaches the load. Since the current on the T-line is $i^+ = v^+/Z_0 = V_s/(Z_0 + R_s)$ and the current at the load is V_L/R_L, a discontinuity is produced at the load.

Thus a reflected wave must be created at discontinuity to satisfy the boundary condition

$$V_L(t) = v^+(\ell,t) + v^-(\ell,t) \tag{3.16}$$

$$I_L(t) = \frac{1}{Z_0}v^+(\ell,t) - \frac{1}{Z_0}v^-(\ell,t) = V_L(t)/R_L \tag{3.17}$$

Solving for the forward and reflected waves

$$2v^+(\ell,t) = V_L(t)(1 + Z_0/R_L) \tag{3.18}$$

$$2v^-(\ell,t) = V_L(t)(1 - Z_0/R_L) \tag{3.19}$$

And therefore the reflection from the load is given by

$$\Gamma_L = \frac{V^-(\ell,t)}{V^+(\ell,t)} = \frac{R_L - Z_0}{R_L + Z_0} \tag{3.20}$$

The reflection coefficient (same as the time harmonic case) plays a very important concept for transmission lines, and for a real termination resistor, $-1 \le \Gamma_L \le 1$. Some noteworthy cases are

- $\Gamma_L = -1$ for $R_L = 0$ (short)
- $\Gamma_L = +1$ for $R_L = \infty$ (open)
- $\Gamma_L = 0$ for $R_L = Z_0$ (match)

Most often, we strive to impedance match to provide the proper termination to avoid reflections. Otherwise if $\Gamma_L \neq 0$, a new reflected wave travels toward the source and unless $R_s = Z_0$, another reflection also occurs at source! To see this consider the wave arriving at the source. Recall that since the wave equation is linear partial differential equation, a superposition of any number of solutions is also a solution. At the source end the boundary condition is as follows

$$V_s - I_s R_s = v_1^+ + v_1^- + v_2^+ \tag{3.21}$$

The new term v_2^+ is used to satisfy the boundary condition. The current continuity requires $I_s = i_1^+ + i_1^- + i_2^+$

$$V_s = (v_1^+ - v_1^- + v_2^+)\frac{R_s}{Z_0} + v_1^+ + v_1^- + v_2^+ \tag{3.22}$$

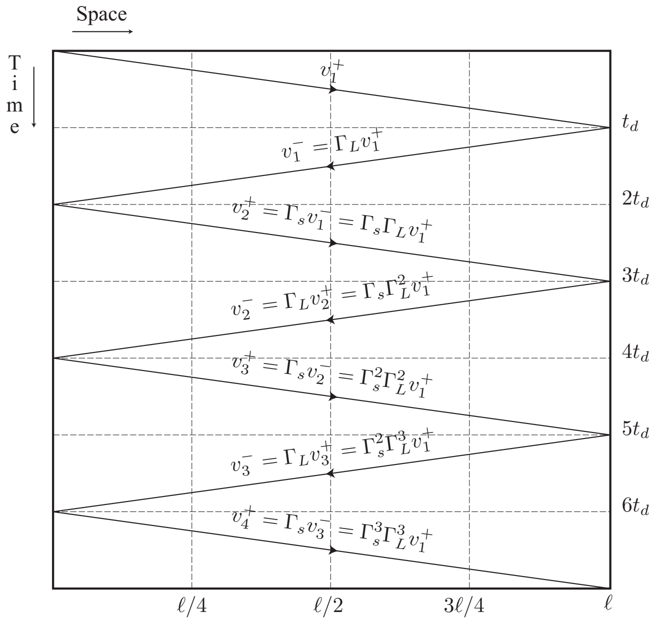

Space
Time
v_1^+
t_d
$v_1^- = \Gamma_L v_1^+$
$2t_d$
$v_2^+ = \Gamma_s v_1^- = \Gamma_s \Gamma_L v_1^+$
$3t_d$
$v_2^- = \Gamma_L v_2^+ = \Gamma_s \Gamma_L^2 v_1^+$
$4t_d$
$v_3^+ = \Gamma_s v_2^- = \Gamma_s^2 \Gamma_L^2 v_1^+$
$5t_d$
$v_3^- = \Gamma_L v_3^+ = \Gamma_s^2 \Gamma_L^3 v_1^+$
$6t_d$
$v_4^+ = \Gamma_s v_3^- = \Gamma_s^3 \Gamma_L^3 v_1^+$
$\ell/4$
$\ell/2$
$3\ell/4$
ℓ

Figure 3.5: The bounce diagram.

Solve for v_2^+ in terms of known terms

$$V_s = \left(1 + \frac{R_s}{Z_0}\right)(v_1^+ + v_2^+) + \left(1 - \frac{R_s}{Z_0}\right) v_1^- \tag{3.23}$$

But $v_1^+ = \frac{Z_0}{R_s+Z_0} V_s$

$$V_s = \frac{R_s + Z_0}{Z_0} \frac{Z_0}{R_s + Z_0} V_s + \left(1 - \frac{R_s}{Z_0}\right) v_1^- + \left(1 + \frac{R_s}{Z_0}\right) v_2^+ \tag{3.24}$$

So the source terms cancel out and

$$v_2^+ = \frac{R_s - Z_0}{Z_0 + R_s} v_1^- = \Gamma_s v_1^- \tag{3.25}$$

The reflected wave bounces off the source impedance with a reflection coefficient given by the same equation as before

$$\Gamma(R) = \frac{R - Z_0}{R + Z_0} \tag{3.26}$$

The source appears as a short for the incoming wave. Invoke superposition! The term v_1^+ took care of the source boundary condition so our new v_2^+ only needed to compensate for the v_1^- wave. The reflected wave is only a function of v_1^-.

The Bounce Diagram

We can track the multiple reflections with a “bounce diagram”, shown in Fig. 3.5. If we freeze time and look at the line, using the bounce diagram we can figure out how many reflections have occurred. For instance, at time $2.5t_d = 2.5\ell/v$ three waves have been excited (v_1^+,v_1^-, v_2^+), but v_2^+ has only traveled a distance of $\ell/2$. To the left of $\ell/2$, the voltage is a summation of three components

$$v = v_1^+ + v_1^- + v_2^+ = v_1^+(1 + \Gamma_L + \Gamma_L \Gamma_s) \tag{3.27}$$

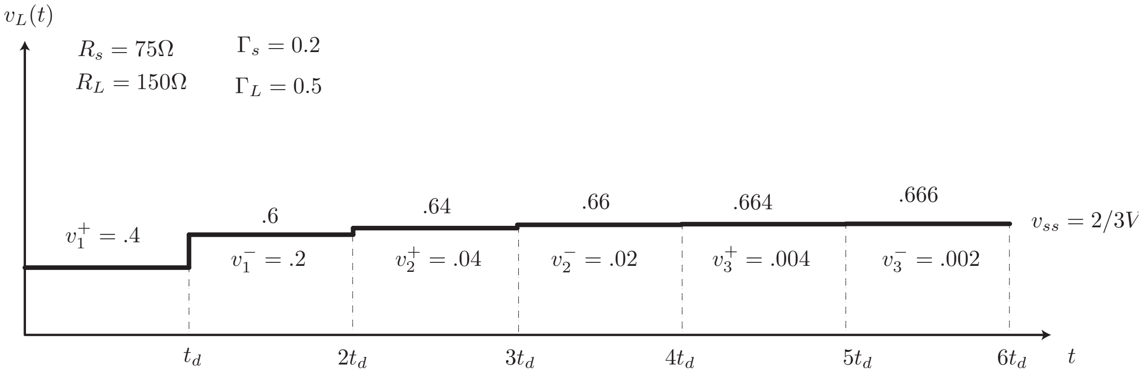


Figure 3.6: The voltage waveform on a distributed transmission line at a given instant of time.

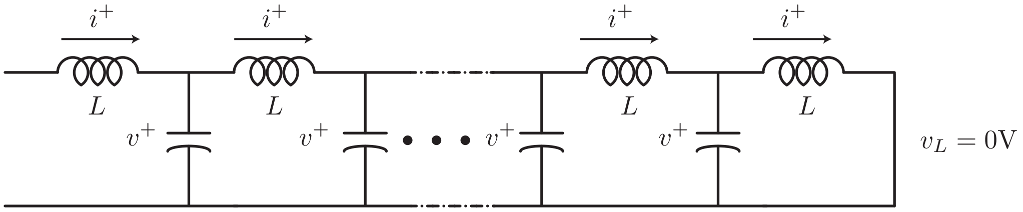


Figure 3.7: An *LC* ladder model of a shorted transmission line.

To the right of $\ell/2$, the voltage has only two components

$$v = v_1^+ + v_1^- = v_1^+(1+\Gamma_L) \tag{3.28}$$

We can also pick at arbitrary point on the line and plot the evolution of voltage as a function of time. For instance, at the load, assuming $R_L > Z_0$ and $R_S > Z_0$, so that $\Gamma_{s,L} > 0$, the voltage at the load will will increase with each new arrival of a reflection, as shown in Fig. 3.6.

Physical Intuition: Shorted Line

To gain physical insight into the transient behavior of a transmission line, consider Fig. 3.7, and *LC* ladder model. The initial step charges the "first" capacitor through the 'first" inductor since the line is uncharged. There is a delay since on the rising edge of the step, the inductor is an open. Each successive capacitor is charged by "its" inductor in a uniform fashion ... this is the forward wave v_1^+.

The voltage on the line goes up from left to right due to the delay in charging each inductor through the inductors. The last inductor, though, does not have a capacitor to charge. Thus the last inductor is "discharged", with extra charge flowing through discharging the "last" capacitor. As this capacitor discharges, so does its neighboring capacitor to the left. Again there is a delay in discharging the caps due to the inductors. This discharging represents the backward wave v_1^-.

Steady-State Waveform

To find steady-state voltage on the line, we sum over all reflected waves

$$v_{ss} = v_1^+ + v_1^- + v_2^+ + v_2^- + v_3^+ + v_3^- + v_4^+ + v_4^- + \cdots \tag{3.29}$$

Or in terms of the first wave on the line

$$v_{ss} = v_1^+(1 + \Gamma_L + \Gamma_L\Gamma_s + \Gamma_L^2\Gamma_s + \Gamma_L^2\Gamma_s^2 + \Gamma_L^3\Gamma_s^2 + \Gamma_L^3\Gamma_s^3 + \cdots \tag{3.30}$$

Notice geometric sums of terms like $\Gamma_L^k\Gamma_s^k$ and $\Gamma_L^{k+1}\Gamma_s^k$. Let $x = \Gamma_L\Gamma_s$

$$v_{ss} = v_1^+(1+x+x^2+\cdots+\Gamma_L(1+x+x^2+\cdots)) \tag{3.31}$$

The sums converge since $|x| < 1$

$$v_{ss} = v_1^+\left(\frac{1}{1-\Gamma_L\Gamma_s} + \frac{\Gamma_L}{1-\Gamma_L\Gamma_s}\right) \tag{3.32}$$

or more compactly

$$v_{ss} = v_1^+\left(\frac{1+\Gamma_L}{1-\Gamma_L\Gamma_s}\right) \tag{3.33}$$

Substituting for Γ_L and Γ_s gives

$$v_{ss} = V_s\frac{R_L}{R_L+R_s} \tag{3.34}$$

In steady state, the equivalent circuit shows that the transmission line has disappeared. This happens because if we wait long enough, the effects of propagation delay do not matter. Conversely, if the propagation speed were infinite, then the T-line would not matter. But the presence of the T-line will be felt if we disconnect the source or load! That's because the T-line *stores* reactive energy in the capacitance and inductance. Note that every real circuit behaves this way! Lumped circuit theory is an abstraction that only applies to infinitesimally small circuits.

Ringing for Open/Short Loads

Consider a source driving an open line shown in Fig. 3.8. Suppose the source impedance is $Z_0/4$, so $\Gamma_s = -0.6$, and the load is open so $\Gamma_L = 1$. As before a positive going wave is launched $v_1^+ = 4V_s/5$. Upon reaching the load, a reflected wave of of equal amplitude is generated and the load voltage overshoots $v_L = v_1^+ + v_1^- = 1.6$V. Note that the current reflection is negative of the voltage

$$\Gamma_i = \frac{i^-}{i^+} = -\frac{v^-}{v^+} = -\Gamma_v \tag{3.35}$$

This means that the sum of the currents at load is zero (open).

At source a new reflection is created $v_2^+ = \Gamma_L\Gamma_s v_1^+$, and note $\Gamma_s < 0$, so $v_2^+ = -.6\times 0.8 = -0.48$. At a time $3t_p$, the line charged initially to $v_1^+ + v_1^-$ drops in value

$$v_L(3t_p) = v_1^+ + v_1^- + v_2^+ + v_2^- = 1.6 - 2\times .48 = .64 \tag{3.36}$$

So the voltage on the line undershoots < 1. And on the next cycle $5t_p$ the load voltage again overshoots. We observe ringing with frequency $2t_p$ as shown in Fig. 3.9. When we drive an unterminated line, especially and open-circuited or a short-circuited line, we observe ringing. In a digital circuit with a long propagation delay t_p, if the ringing is severe enough, it may require us to slow down the clock to allow the voltage to properly settle. This is one of the main advantage of a terminated transmission line in such a system.

Transmission Line Resonator

In Chapter 2 we found that transmission lines resonate when the electrical length becomes a multiple of the quarter wavelength. At odd multiples the line behaves like a parallel *RLC* circuit whereas at even multiples it behaves like a series *RLC* circuit. In the time domain, we shall see

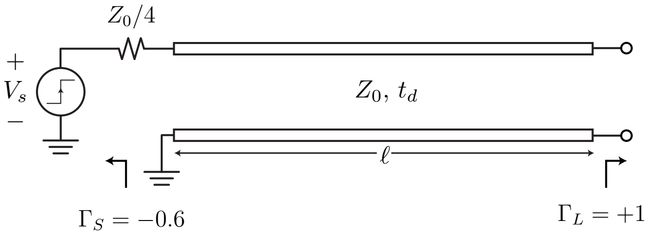


Figure 3.8: A step function voltage source drives an open line.

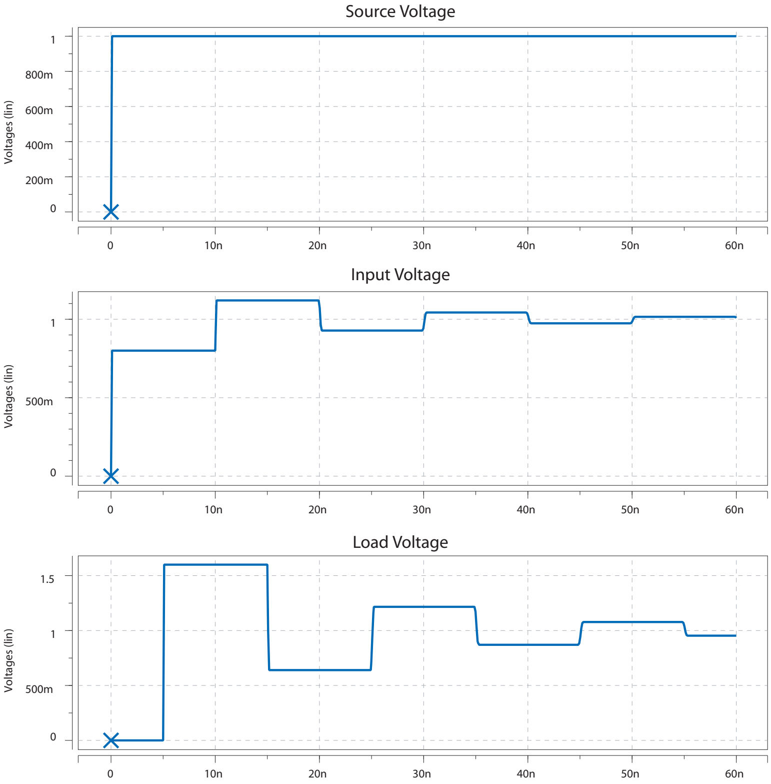


Figure 3.9: SPICE simulation shows ringing waveforms for the transmission line shown in Fig. 3.8. The time delay of the transmission line is $t_p = 5\,\text{ns}$.

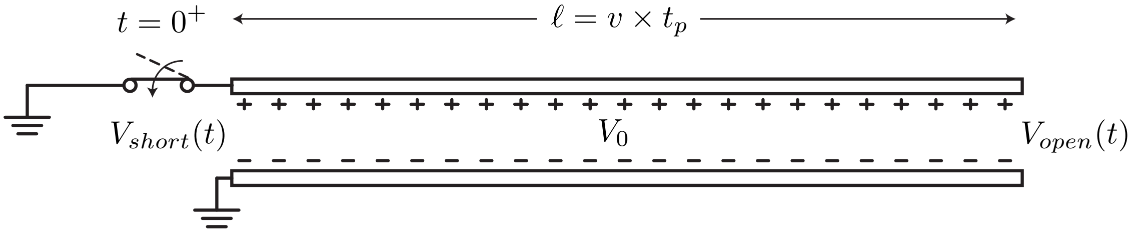


Figure 3.10: A transmission line initially charged to a voltage V_0 is discharged by closing the switch at time $t = 0^+$.

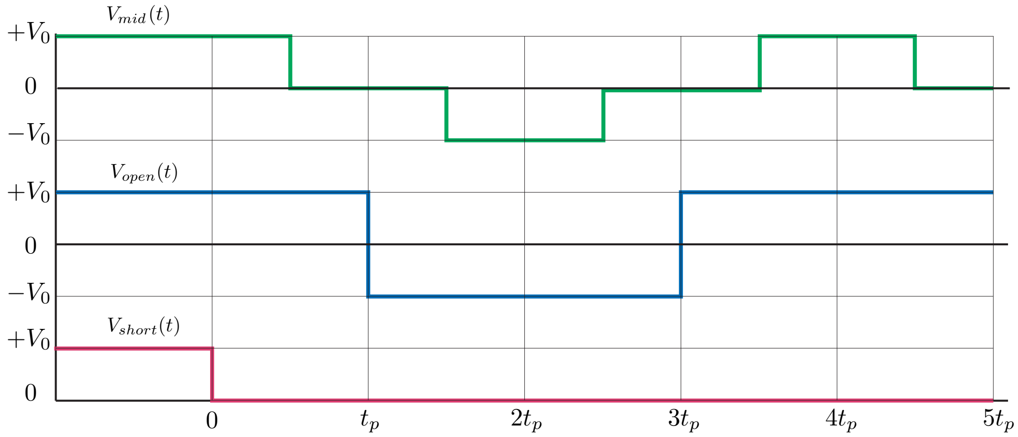


Figure 3.11: The voltage waveforms on the initially charged transmission line at points $z = 0$ (shorted end), $z = \ell$ (open end), and at the mid-point.

that a transmission line can be used to generate periodic waveforms with fast transitions and short periods.

The setup is shown in Fig. 3.10. We assume that the transmission is initially charged to a voltage V_0 and then at time $t = 0^+$ we short the transmission line to ground. Since the transmission line is lossless, the energy stored in the line, $CV^2/2$, cannot disappear. Instead as the line is shorted, it will be converted into magnetic energy $LI^2/2$, and then back again to electrical energy. So we see a resonance occurring on the line but instead of sinusoidal waves, the line will have a square wave voltage and current waveform. From a frequency domain perspective, we see that the initial closure of the switch excites multiple harmonics on the transmission line which sum to form a square wave. Since the initial step is rich in harmonics, it will in theory impart energy onto an infinite number of harmonics frequencies.

Let's solve this problem in the following manner. At time $t = 0$, the voltage on the line is given by V_0. When the switch is closed, we generate a wave $V_1^+ = -V_0$ at the position of the switch which discharges the voltage there. This voltage wave travels towards the open end of the transmission line at the speed of light in the medium, $v = \sqrt{1/L'C'}$, and as it travels it discharges the voltage on the line to zero. In a time $t = t_p = \ell/v$, this wave will reach the open end and the voltage on the entire line is reduced to zero. Now all the energy of the line is contained in the current, or the magnetic energy. Due to the open boundary condition, though, the voltage V^+ is reflected and now a new wave $V_1^- = V_1^+ = -V_0$ is generated. This wave now travels leftward towards the shorted end of the transmission line. As it travels, it reduces the voltage on the line to $-V_0$. Now we can see that this voltage, upon reaching the shorted end, will generate a new voltage, $V_2^+ = -V_2^- = V_0$, which travels forward towards the open end. At the shorted end of the voltage each subsequent wave generated always cancels the incoming voltage to produce a net zero voltage.

In this way we see that the voltage on the transmission line oscillates between V_0 and $-V_0$ as a function of time at the load. At the short end it drops to zero as the switch is closed. At the open end, though, the voltage will oscillate. The period of this oscillation is given by the time delay of the transmission line, or $T = 2t_p$. A plot of the voltage at the open and short end is shown in Fig. 3.11. Interestingly, at the midpoint of the transmission line, the voltage waveform is more complicated as partial reflections drive the line to zero volts, then negative, and the positive again. Since the line is lossless, this oscillation will ensue until the line is somehow disturbed again. Very fast oscillations can be produced in this way. Unfortunately every real transmission line has loss and so the voltage waveform will decay towards zero. If we can somehow introduce energy back

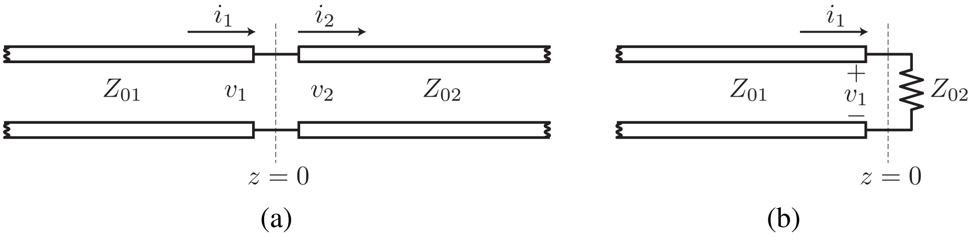


Figure 3.12: (a) The cascade of two transmission lines. (b) The equivalent load presented by the second transmission line.

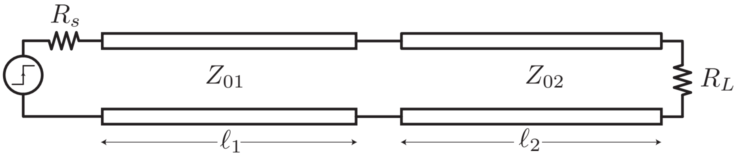


Figure 3.13: The load is connected to the source by a cascade of two sections of transmission lines.

onto the line, we can in fact produce a distributed oscillator.

Cascade of Transmission Lines

Consider the junction between two transmission lines Z_{01} and Z_{02}, shown in Fig. 3.12a. At the interface $z=0$, the boundary conditions are that the voltage/current have to be continuous

$$v_1^+ + v_1^- = v_2^+ \tag{3.37}$$

$$(v_1^+ - v_1^-)/Z_{01} = v_2^+/Z_{02} \tag{3.38}$$

Solve these equations in terms of v_1^+. The reflection coefficient has the same form (easy to remember)

$$\Gamma = \frac{v_1^-}{v_1^+} = \frac{Z_{02}-Z_{01}}{Z_{01}+Z_{02}} \tag{3.39}$$

As shown in Fig. 3.12b, the second line looks like a load impedance of value Z_{02}. The wave launched on the new transmission line at the interface is given by

$$v_2^+ = v_1^+ + v_1^- = v_1^+(1+\Gamma) = \tau v_1^+ \tag{3.40}$$

This "transmitted" wave has a coefficient

$$\tau = 1+\Gamma = \frac{2Z_{02}}{Z_{01}+Z_{02}} \tag{3.41}$$

Notice that the sum of τ^2 and Γ^2 is unity

$$\tau^2 + \Gamma^2 = 1 \tag{3.42}$$

This follows from conservation of energy. We can construct a bounce diagram for the scenario shown in Fig. 3.13. Thus reflections occur at multiple interfaces, shown in Fig. 3.14.

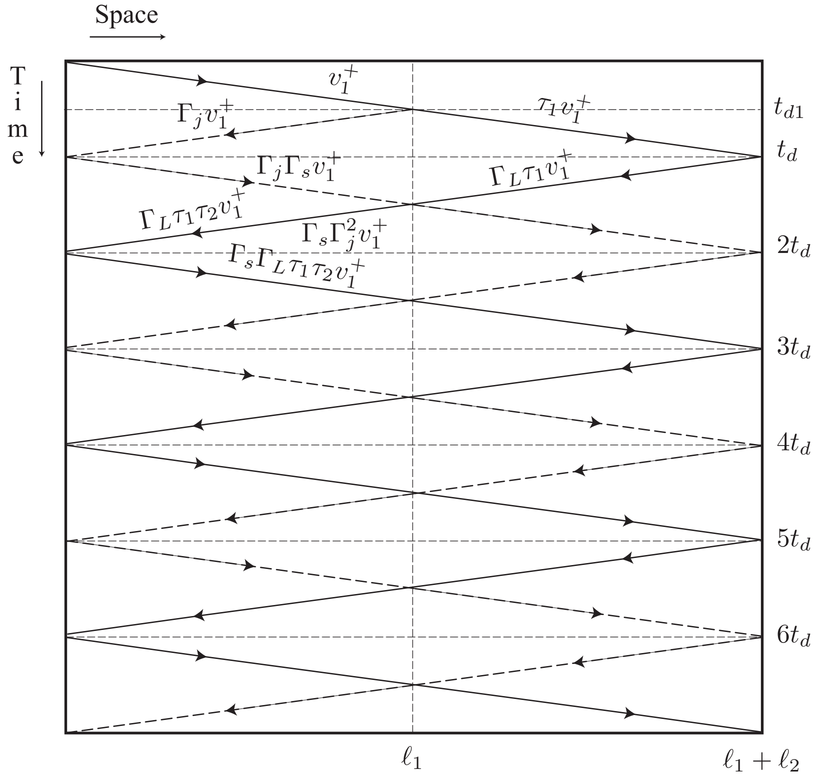


Figure 3.14: Bounce diagram for the setup shown in Fig. 3.13.

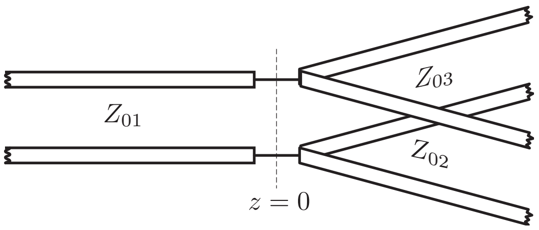


Figure 3.15: A transmission line drives two parallel lines.

Junction of Parallel T-Lines

Consider the junction of three transmission lines, shown in Fig. 3.15. Using the same approach as before, we invoke voltage/current continuity at the interface

$$v_1^+ + v_1^- = v_2^+ = v_3^+ \tag{3.43}$$

$$\frac{v_1^+ - v_1^-}{Z_{01}} = \frac{v_2^+}{Z_{02}} + \frac{v_3^+}{Z_{03}} \tag{3.44}$$

But $v_2^+ = v_3^+$, so the interface just looks like the case of the junction of two transmission lines, Z_{01} and a new line with characteristic impedance $Z_{01}||Z_{02}$.

Example 1:

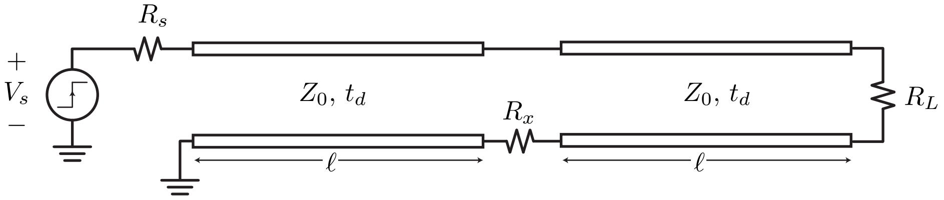


Figure 3.16: A layout error introduces a ground resistance between two transmission lines.

Consider the following circuit consisting of two transmission lines of characteristic impedance Z_0. Due to a layout error, the ground connection is not good and presents a resistance of $R_x = .5Z_0$. The circuit is excited by a pulse at the generator with amplitude 1V and source impedance $R_s = 2Z_0$ (zero rise-time). The load is matched $R_L = Z_0$.

We analyze this problem by drawing the bounce diagram, shown in Fig. 3.17, for the circuit. We only include the action on the first transmission line. From the bounce diagram, we can sketch the voltage waveform at the load and at an arbitrary point, such s $z/\ell = .25$ as a function of time on the first transmission line. These graphs are shown in Fig. 3.18.

3.6 Reactive Terminations

Consider a reactive termination. This situation is common in practice because a shorted or open load has non-zero reactance. Let's analyze the problem of an inductive load first. When a pulse first "sees" the inductance at the load, it looks like an open so $\Gamma_0 = +1$. As time progresses, the inductor looks more and more like a short! So $\Gamma_\infty = -1$.

So intuitively we might expect the reflection coefficient to look like Fig. 3.19. The graph starts at $+1$ and ends at -1. In between we'll see that it goes through exponential decay (1st order ODE).

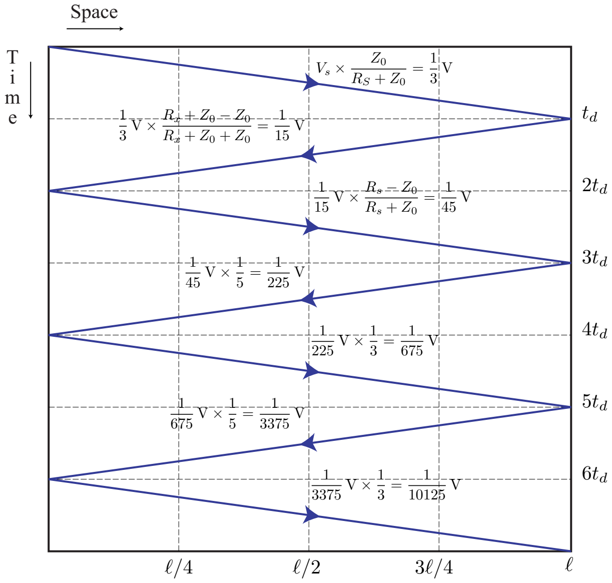
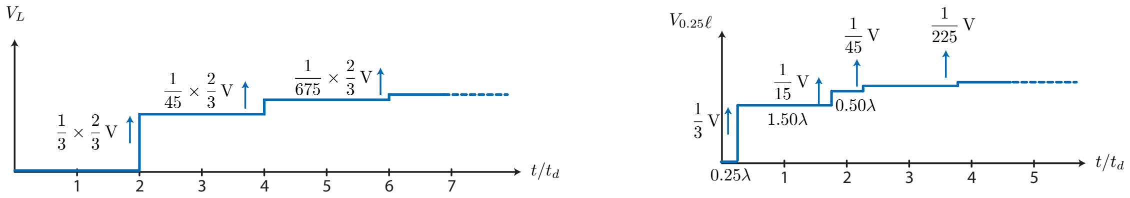

Space
Time
$V_s \times \frac{Z_0}{R_S + Z_0} = \frac{1}{3}$V
$\frac{1}{3}$V $\times \frac{R_x + Z_0 - Z_0}{R_x + Z_0 + Z_0} = \frac{1}{15}$V
$\frac{1}{15}$V $\times \frac{R_s - Z_0}{R_s + Z_0} = \frac{1}{45}$V
$\frac{1}{45}$V $\times \frac{1}{5} = \frac{1}{225}$V
$\frac{1}{225}$V $\times \frac{1}{3} = \frac{1}{675}$V
$\frac{1}{675}$V $\times \frac{1}{5} = \frac{1}{3375}$V
$\frac{1}{3375}$V $\times \frac{1}{3} = \frac{1}{10125}$V
t_d
$2t_d$
$3t_d$
$4t_d$
$5t_d$
$6t_d$
$\ell/4$
$\ell/2$
$3\ell/4$
ℓ

Figure 3.17: Bounce diagram for discontinuity shown in Fig. 3.16.



(a) (b)

Figure 3.18: (a) The voltage variation at the load. (b) The voltage variation at a distance of 0.25λ on the first transmission line. Note drawings are not drawn to scale (vertical).

Do equations confirm our intuition?

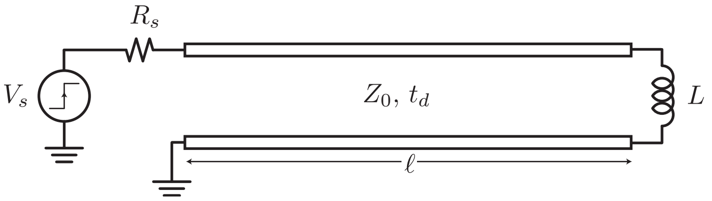
$$v_L = L\frac{di}{dt} = L\frac{d}{dt}\left(\frac{v^+}{Z_0} - \frac{v^-}{Z_0}\right) \tag{3.45}$$

And the voltage at the load is given by $v^+ + v^-$

$$v^- + \frac{L}{Z_0}\frac{dv^-}{dt} = \frac{L}{Z_0}\frac{dv^+}{dt} - v^+ \tag{3.46}$$

The right hand side is known, it's the incoming waveform. For the step response, the derivative term on the RHS is zero at the load

$$v^+ = \frac{Z_0}{Z_0 + R_s}V_s \tag{3.47}$$



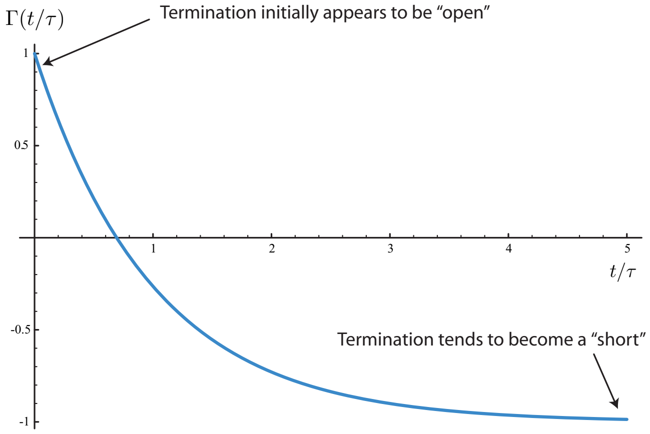


Figure 3.19: The reflection coefficient versus time for an inductive load.

So we have a simpler case $\frac{dv^+}{dt} = 0$. We must solve the following equation

$$v^- + \frac{L}{Z_0}\frac{dv^-}{dt} = -v^+ \tag{3.48}$$

For simplicity, assume at $t = 0$ the wave v^+ arrives at load. In the Laplace domain

$$V^-(s) + \frac{sL}{Z_0}V^-(s) - \frac{L}{Z_0}v^-(0) = -v^+/s \tag{3.49}$$

Solve for reflection $V^-(s)$

$$V^-(s) = \frac{v^-(0)L/Z_0}{1+sL/Z_0} - \frac{v^+}{s(1+sL/Z_0)} \tag{3.50}$$

Break this into basic terms using partial fraction expansion

$$\frac{-1}{s(1+sL/Z_0)} = \frac{-1}{s} + \frac{L/Z_0}{1+sL/Z_0} \tag{3.51}$$

Invert the equations to get back to time domain $t > 0$

$$v^-(t) = (v^-(0) + v^+)e^{-t/\tau} - v^+ \tag{3.52}$$

Note that $v^-(0) = v^+$, since initially the inductor is an open. So the reflection coefficient is

$$\Gamma(t) = 2e^{-t/\tau} - 1 \tag{3.53}$$

The reflection coefficient decays with time constant L/Z_0

In fact, we can simplify the above procedure by using Laplace analysis direction. We can find the reflection coefficient $\Gamma(s)$ directly in the Laplace domain without any equations. For a periodic excitation, such as a square wave, we can use Fourier analysis directly. This is illustrated by the next example.

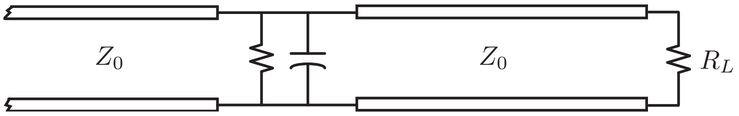


Figure 3.20: A shunt *RC* discontinuity.

Example 2:

The transmission line shown in Fig. 3.20 has a discontinuity connected in shunt as shown below. The discontinuity consists of a capacitor and a parallel resistor. The load is matched to the line. Derive the equations governing the reflection v^- at the discontinuity.

The load seen by the first transmission line is given by $R||C||Z_0$, or

$$Z_L = \left(\frac{1}{R} + \frac{1}{Z_0} + sC\right)^{-1}$$

$$= \frac{RZ_0}{R + Z_0 + sRZ_0C}$$

The complex reflection coefficient seen by the first transmission line is given by

$$\Gamma(s) = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{-(Z_0^2 + RZ_0^2 sC)}{2RZ_0 + Z_0^2 + RZ_0^2 sC}$$

We can find $V^- = \rho V^+$ by using Fourier or Laplace transform analysis.

3.7 Transmission Line Dispersion

In reality all transmission lines deviate from ideal behavior in two important ways. First, all transmission lines have loss, and thus a wave on a transmission line decays in magnitude as it travels down the line. Second, most real transmission lines have non-linear phase delay versus frequency, leading to waveform dispersion. In Chapter 2 we found that we could quantify the loss in terms of the real part $\alpha(\omega)$ of the complex propagation constant $\gamma = \alpha + j\beta$ and the phase constant is determined by the imaginary component $\beta(\omega)$.

Signal decay can be compensated by placing amplifiers, or repeaters, along the transmission line. But the waveforms do not shrink by simple scaling as magnitude distortion occurs since α is not a constant function of frequency. The different frequency components of the waveform experience a different attenuation, with the attenuation typically an increasing function of frequency. The high frequency components of a signal, such as the sharp edges, tend to be softened, as they travel along a transmission line.

Frequency dependent losses occur for multiple reasons. In most transmission line structures, for instance, the conductive losses increase like $\sqrt{f}$ due to skin-effect. Other loss mechanisms, such as dielectric loss, also increase with frequency, due to dielectric resonance.

In theory we can compensate for the frequency dependent losses by introducing the correct compensation network. For instance, an amplifier can boost high frequencies (with a zeros in the transfer function) to clean up the waveform. But in a typical transmission line we have to also compensate for the waveform distortion due to the non-linear phase constant $\beta(\omega)$. Recall that a non-distorting transfer function should ideally scale and delay a waveform. An ideal transmission line has this ideal characteristic as the phase constant is a linear function of frequency. For a lossless transmission line we have

$$\gamma = \sqrt{j\omega L j\omega C} = j\omega\sqrt{LC} \tag{3.54}$$

a linear relation between the phase delay versus frequency. The group delay for any waveform is therefore a constant. Since switching waveforms are broadband in nature, we need to ensure that over the bandwidth of the waveform, the phase delay has a linear response. Intuitively it's clear that a non-uniform group delay implies that different parts of the waveform arrive at different times and thus corrupt the waveform. For a lossy transmission line where the losses are dominated by the conductive losses of the material, we have

$$\gamma = \sqrt{(j\omega L + R)j\omega C} = j\omega\sqrt{LC}\sqrt{1 - j\frac{R}{\omega L}} \tag{3.55}$$

For low frequencies, or if $\omega L \ll R$, the line is an RC distributed line that we studied in Chapter **??**. For this case the attenuation and group delay is frequency dependent leading to distortion. For high frequencies, though, $\omega \gg R/L$, the line behaves like an ideal low-loss transmission line. Unfortunately for a real line the conductive losses increase with frequency and lead dispersion. Furthermore, inclusion of dielectric losses leads to further distortion.

To demonstrate the effects of dispersion, consider the following wave packet traveling on a transmission line

$$p(x,t) = \sum_{k=1}^{100} 0.07\left(e^{-(0.1k-3)^2} + e^{-(0.1k+3)^2}\right)\cos\left(0.1kx - \frac{0.1kt}{\sqrt{1+0.02(0.1k)^2}}\right) \tag{3.56}$$

The waveform dispersion occurs since the phase delay decreases non-linearly for each successive harmonic k of the signal, as shown in Fig. 3.21. A plot of the waveform shape traveling down the line is shown in Fig. 3.22. In Fig. 3.22a we have removed the dispersion to show the an ideal case of a wavepacket traveling to the right. In Fig. 3.22b we see severe distortion occurring. Since higher frequencies travel slower, the wave packet broadens and distorts along the line.

We found from Eq. 5.71 that if $R/L = G/C$, then the lossy transmission line is also dispersionless. This is why old telephone networks included artificial transmission lines with lumped inductor sections, used to raise the inductance per unit length to satisfy this equation. In a modern communication system we solve this problem using the power of digital electronics. In essence, if we could discover the phase and magnitude response of the channel, the transmission line in this case, then we could construct a digital filter to compensate for the phase non-linearity of the channel. This is known as equalization. Alternatively, we can break our signals into multiple narrowband carriers, and send these carriers through the channel. Since each carrier is narrowband, phase distortion does not occur since over a narrow bandwidth the group delay is approximately constant. Inter-carrier interference can be minimized by using orthogonal carrier tones. We can also distribute energy into the tones in such a way to maximize the overall performance of the system. Since different tones will experience different levels of attenuation, it's intuitively clear that we should inject more power into weak tones to maximize the overall signal to noise ratio. This is in fact the manner in which modern communication systems compensate for the channel properties. Digital subscriber line (DSL) uses an approach similar to this to pack significantly more data onto a telephone wire than otherwise possible.[2] A similar scheme is employed in OFDM broadband wireless systems such as 802.11a/g wireless LAN.

[2] Recall the days of a 300 baud modem? Well at least you recall the 56K modem. A DSL link can pump 1.5 Mbps through the same line.

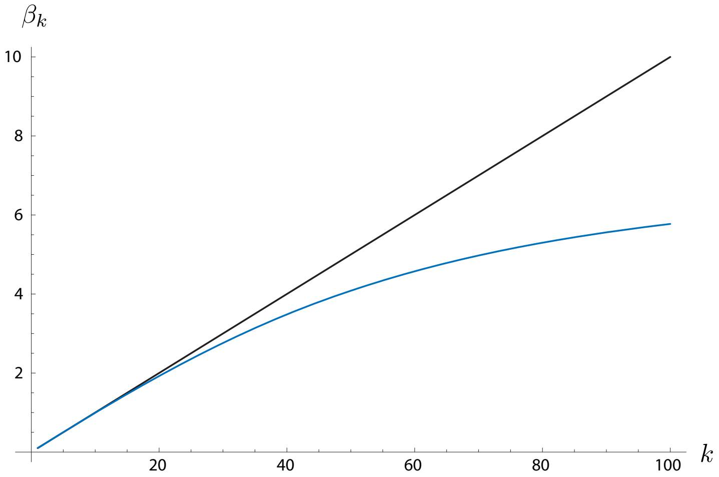


Figure 3.21: The dispersion curve shows that increasing harmonics k of the wavepacket have smaller propagation constants β_k.

3.8 References

This chapter has been adapted from my lecture notes prepared for an electromagnetics course. I have used several references, including *Fields and Waves in Communication Electronics* [**ramo**], Cheng [**cheng**], and Inan and Inan [**inan**]. Other important references include Johnson and Graham [**magic**].

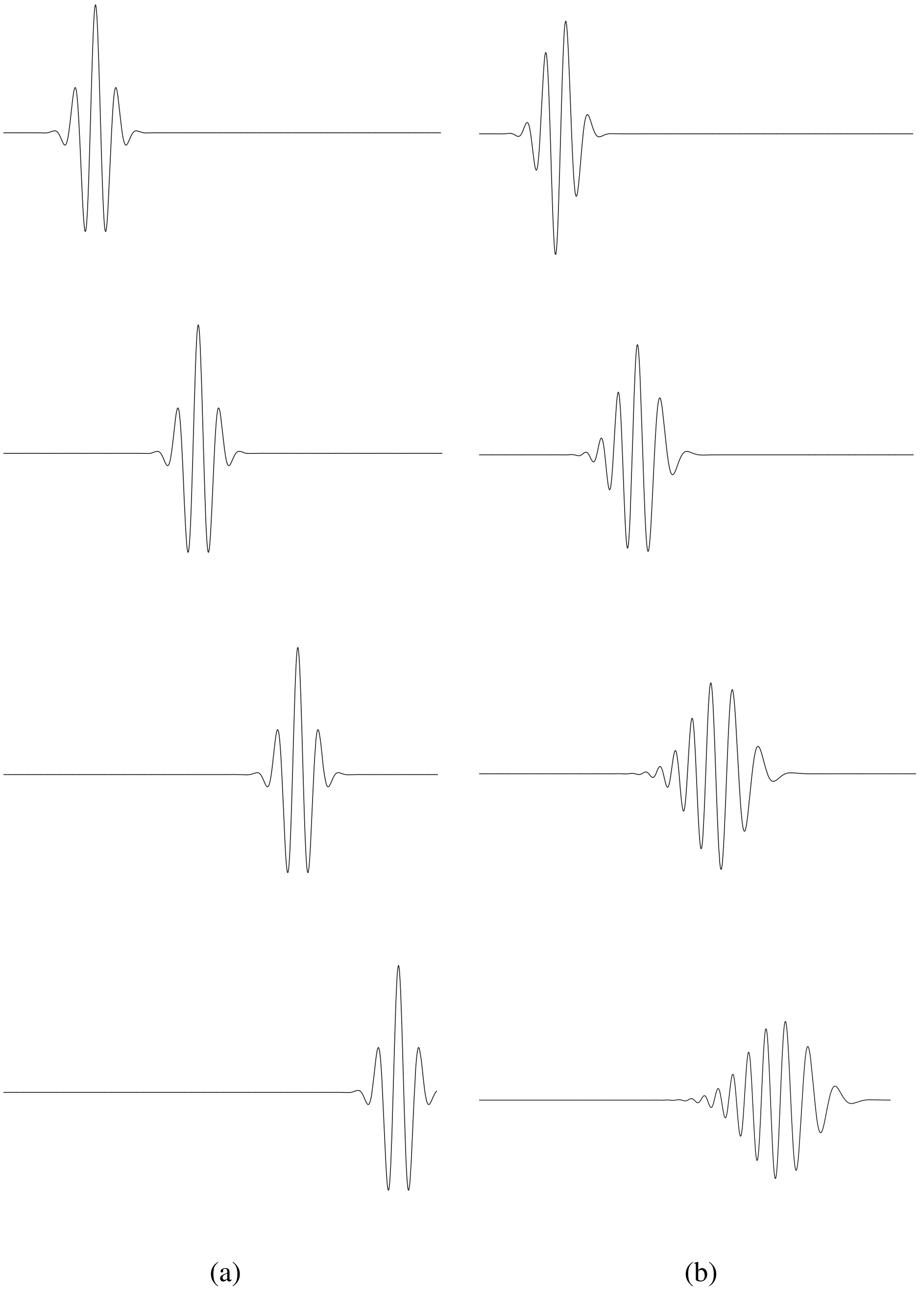
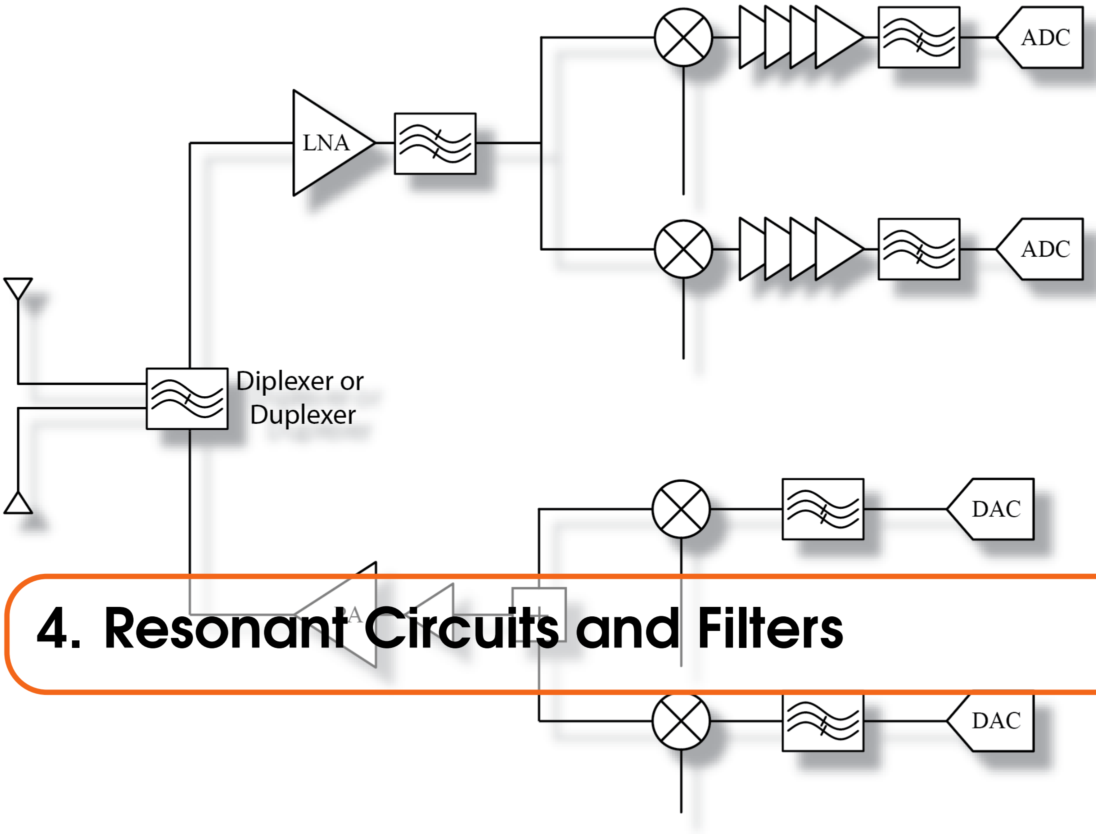


Figure 3.22: (a) A wave packet travels at constant speed along a transmission line with no dispersion. (b) A wave packet experiences severe dispersion as it travels through a transmission line.



4. Resonant Circuits and Filters

4.1 Introduction

In this chapter we will study, design, and "build" *LC* low-pass and band-pass filters. Filters are essential components in communication systems, allowing many users to share the spectrum through frequency division multiplexing. Closely spaced channels allow efficient spectrum utilization but require high selectivity filtering to extract wanted signals from a background of noise and interference.

In the appendix we will introduce basic surface mount components and effective soldering techniques. The non-ideal effects of the components, such as loss and self-resonance, as well as the board parasitic inductance, capacitance, and delay, will be highlighted.

4.2 Resonance

Because *RLC* circuits and resonance play such a crucial in filters, we begin with a review of this topic. These circuits are simple enough to allow full analysis, and yet rich enough to form the basis for most of the circuits we will study.

Series *RLC* Circuits

The *RLC* circuit shown in Fig. 4.1 is deceptively simple. The impedance seen by the source is simply given by

$$Z = j\omega L + \frac{1}{j\omega C} + R = R + j\omega L\left(1 - \frac{1}{\omega^2 LC}\right) \tag{4.1}$$

The impedance is purely real at at the *resonant frequency* when $\Im(Z) = 0$, or $\omega = \pm\frac{1}{\sqrt{LC}}$. At resonance the impedance takes on a minimal value. It's worthwhile to investigate the cause of resonance, or the cancellation of the reactive components due to the inductor and capacitor. Since the inductor and capacitor voltages are always 180° out of phase, and one reactance is dropping while the other is increasing, there is clearly always a frequency when the magnitudes are equal. Thus resonance occurs when $\omega L = \frac{1}{\omega C}$. A phasor diagram, shown in Fig. 4.2, shows this in detail.

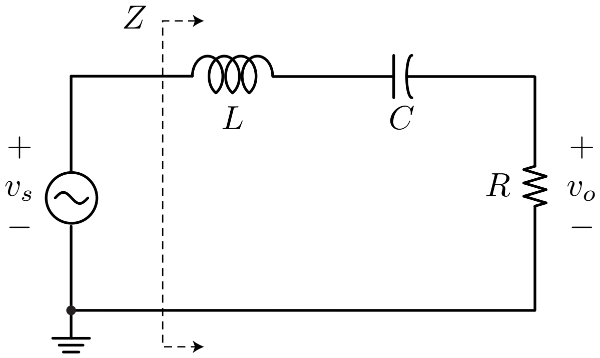


Figure 4.1: A series *RLC* circuit.

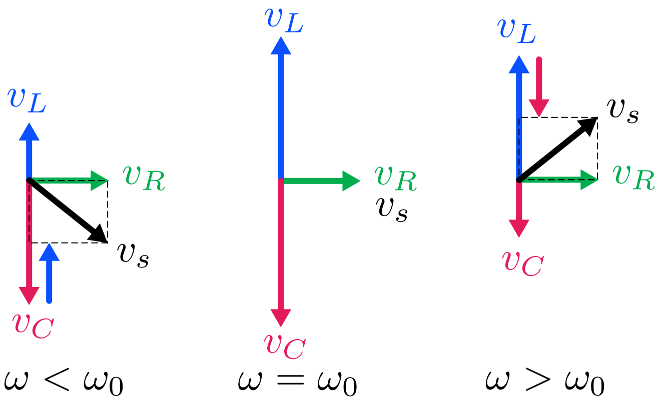


Figure 4.2: The phasor diagram of voltages in the series *RLC* circuit (a) below resonance, (b) at resonance, and (c) beyond resonance.

So what's the magic about this circuit? The first observation is that at resonance, the voltage across the reactances can be larger, in fact much larger, than the voltage across the resistors R. In other words, this circuit has voltage gain. Of course it does not have power gain, for it is a passive circuit. The voltage across the inductor is given by

$$v_L = j\omega_0 L i = j\omega_0 L \frac{v_s}{Z(j\omega_0)} = j\omega_0 L \frac{v_s}{R} = jQ \times v_s \tag{4.2}$$

where we have defined a circuit Q factor at resonance as

$$Q = \frac{\omega_0 L}{R} \tag{4.3}$$

It's easy to show that the same voltage multiplication occurs across the capacitor

$$v_C = \frac{1}{j\omega_0 C} i = \frac{1}{j\omega_0 C} \frac{v_s}{Z(j\omega_0)} = \frac{1}{j\omega_0 C} \frac{v_s}{R} = -jQ \times v_s \tag{4.4}$$

This voltage multiplication property is the key feature of the circuit that allows it to be used as an impedance transformer.

It's important to distinguish this Q factor from the intrinsic Q of the inductor and capacitor. For now, we assume the inductor and capacitor are ideal. We can re-write the Q factor in several equivalent forms owing to the equality of the reactances at resonance

$$Q = \frac{\omega_0 L}{R} = \frac{1}{\omega_0 C} \frac{1}{R} = \frac{\sqrt{LC}}{C} \frac{1}{R} = \sqrt{\frac{L}{C}} \frac{1}{R} = \frac{Z_0}{R} \tag{4.5}$$

where we have defined the $Z_0 = \sqrt{\frac{L}{C}}$ as the characteristic impedance of the circuit.

4.2.1 Circuit Transfer Function

Let's now examine the transfer function of the circuit

$$H(j\omega) = \frac{v_o}{v_s} = \frac{R}{j\omega L + \frac{1}{j\omega C} + R} \tag{4.6}$$

$$H(j\omega) = \frac{j\omega RC}{1 - \omega^2 LC + j\omega RC} \tag{4.7}$$

Obviously, the circuit cannot conduct DC current, so there is a zero in the transfer function. The denominator is a quadratic polynomial. It's worthwhile to put it into a standard form that quickly reveals important circuit parameters

$$H(j\omega) = \frac{j\omega \frac{R}{L}}{\frac{1}{LC} + (j\omega)^2 + j\omega \frac{R}{L}} \tag{4.8}$$

Using the definition of Q and ω_0 for the circuit

$$H(j\omega) = \frac{j\omega \frac{\omega_0}{Q}}{\omega_0^2 + (j\omega)^2 + j\frac{\omega\omega_0}{Q}} \tag{4.9}$$

Factoring the denominator with the assumption that $Q > \frac{1}{2}$ gives us the complex poles of the circuit

$$s^{\pm} = -\frac{\omega_0}{2Q} \pm j\omega_0 \sqrt{1 - \frac{1}{4Q^2}} \tag{4.10}$$

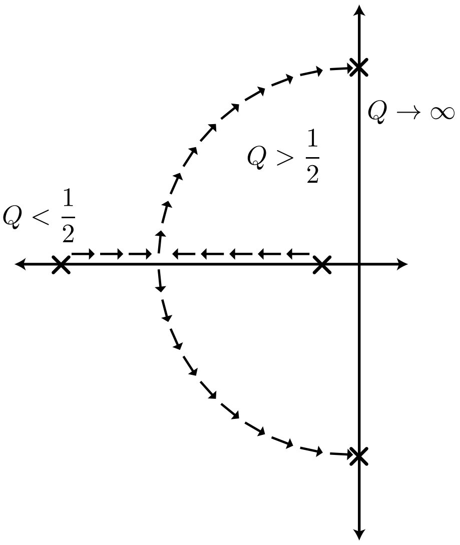


Figure 4.3: The root locus of the poles for a second-order transfer function as a function of Q. The poles begin on the real axis for $Q < \frac{1}{2}$ and become complex, tracing a semi-circle for increasing Q.

The poles have a constant magnitude equal to the resonant frequency

$$|s| = \sqrt{\frac{\omega_0^2}{4\cancel{Q^2}} + \omega_0^2\left(1 - \frac{1}{4\cancel{Q^2}}\right)} = \omega_0 \tag{4.11}$$

A root-locus plot of the poles as a function of Q appears in Fig. 4.3. As $Q \to \infty$, the poles move to the imaginary axis. In fact, the real part of the poles is inversely related to the Q factor.

4.2.2 Circuit Bandwidth

As shown in Fig. 4.4, when we plot the magnitude of the transfer function, we see that the selectivity of the circuit is also related inversely to the Q factor. In the limit that $Q \to \infty$, the circuit is infinitely selective and only allows signals at resonance ω_0 to travel to the load. Note that the peak gain in the circuit is always unity, regardless of Q, since at resonance the L and C together disappear and effectively all the source voltage appears across the load.

The selectivity of the circuit lends itself well to filter applications. To characterize the peakiness, let's compute the frequency when the magnitude squared of the transfer function drops by half

$$|H(j\omega)|^2 = \frac{\left(\omega\frac{\omega_0}{Q}\right)^2}{\left(\omega_0^2 - \omega^2\right)^2 + \left(\omega\frac{\omega_0}{Q}\right)^2} = \frac{1}{2} \tag{4.12}$$

This happens when

$$\left(\frac{\omega_0^2 - \omega^2}{\omega_0\omega/Q}\right)^2 = 1 \tag{4.13}$$

Solving the above equation yields four solutions, corresponding to two positive and two negative frequencies. The peakiness is characterized by the difference between these frequencies, or the bandwidth, given by

$$\Delta\omega = \omega_+ - \omega_- = \frac{\omega_0}{Q} \tag{4.14}$$

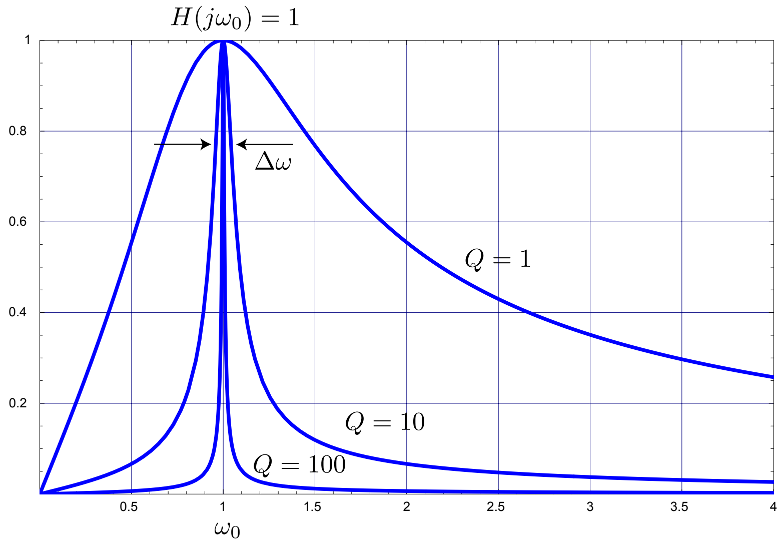


Figure 4.4: The transfer function of a series *RLC* circuit. The output voltage is taken at the resistor terminals. Increasing Q leads to a more peaky response.

which shows that the normalized bandwidth is inversely proportional to the circuit Q

$$\frac{\Delta\omega}{\omega_0} = \frac{1}{Q} \tag{4.15}$$

You can also show that the resonance frequency is the geometric mean frequency of the 3 dB frequencies

$$\omega_0 = \sqrt{\omega_+ \omega_-} \tag{4.16}$$

4.2.3 Energy Storage in *RLC* "Tank"

Let's compute the ebb and flow of the energy at resonance. To begin, let's assume that there is negligible loss in the circuit. The energy across the inductor is given by

$$w_L = \tfrac{1}{2}Li^2(t) = \tfrac{1}{2}LI_M^2 \cos^2 \omega_0 t \tag{4.17}$$

Likewise, the energy stored in the capacitor is given by

$$w_C = \tfrac{1}{2}Cv_C^2(t) = \tfrac{1}{2}C\left(\frac{1}{C}\int i(\tau)d\tau\right)^2 \tag{4.18}$$

Performing the integral leads to

$$w_C = \frac{1}{2}\frac{I_M^2}{\omega_0^2 C}\sin^2 \omega_0 t \tag{4.19}$$

The total energy *stored* in the circuit is the sum of these terms

$$w_s = w_L + w_C = \tfrac{1}{2}I_M^2\left(L\cos^2 \omega_0 t + \frac{1}{\omega_0^2 C}\sin^2 \omega_0 t\right) = \tfrac{1}{2}I_M^2 L \tag{4.20}$$

which is a constant! This means that the reactive stored energy in the circuit does not change and simply moves between capacitive energy and inductive energy. When the current is maximum across the inductor, all the energy is in fact stored in the inductor

$$w_{L,\max} = w_s = \tfrac{1}{2}I_M^2 L \tag{4.21}$$

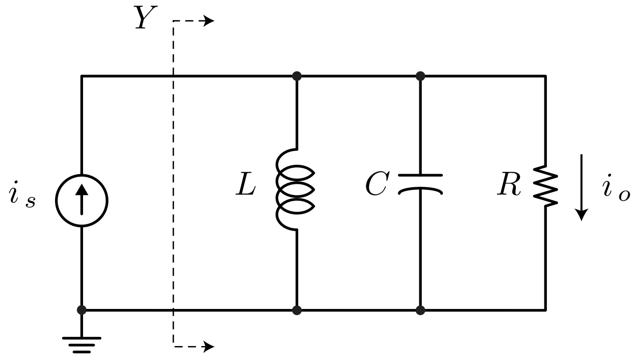


Figure 4.5: A parallel *RLC* circuit.

Likewise, the peak energy in the capacitor occurs when the current in the circuit drops to zero

$$w_{C,\max} = w_s = \tfrac{1}{2}V_M^2 C \tag{4.22}$$

Now let's re-introduce loss in the circuit. In each cycle, a resistor R will dissipate energy

$$w_d = P \cdot T = \tfrac{1}{2}I_M^2 R \cdot \frac{2\pi}{\omega_0} \tag{4.23}$$

The ratio of the energy stored to the energy dissipated is thus

$$\frac{w_s}{w_d} = \frac{\frac{1}{2}LI_M^2}{\frac{1}{2}I_M^2 R\frac{2\pi}{\omega_0}} = \frac{\omega_0 L}{R}\frac{1}{2\pi} = \frac{Q}{2\pi} \tag{4.24}$$

This gives us the physical interpretation of the *Quality Factor* Q as 2π times the ratio of energy stored per cycle to energy dissipated per cycle in an *RLC* circuit

$$Q = 2\pi\frac{w_s}{w_d} \tag{4.25}$$

We can now see that if $Q \gg 1$, then an initial energy in the tank tends to slosh back and forth for many cycles. In fact, we can see that in roughly Q cycles, the energy of the tank is depleted.

4.2.4 Parallel *RLC* Circuits

The parallel *RLC* circuit shown in Fig. 4.5 is the dual of the series circuit. By "dual" we mean that the role of voltage and current are interchanged. Hence the circuit is most naturally probed with a current source i_s. In other words, the circuit has current gain as opposed to voltage gain, and the admittance minimizes at resonance as opposed to the impedance. Finally, the role of capacitance and inductance are also interchanged. In principle, therefore, we don't have to repeat all the detailed calculations we just performed for the series case, but in practice it's worthwhile exercise.

The admittance of the circuit is given by

$$Y = j\omega C + \frac{1}{j\omega L} + G = G + j\omega C\left(1 - \frac{1}{\omega^2 LC}\right) \tag{4.26}$$

which has the same form as Eq. 4.1. The resonant frequency also occurs when $\Im(Y) = 0$, or when $\omega = \omega_0 = \pm\frac{1}{\sqrt{LC}}$. Likewise, at resonance the admittance takes on a minimal value. Equivalently, the impedance at resonance is maximum. This property makes the parallel *RLC* circuit an important element in tuned amplifier loads. It's also easy to show that at resonance the circuit has a current gain of Q

$$i_C = j\omega_0 C v_o = j\omega_0 C \frac{i_s}{Y(j\omega_0)} = j\omega_0 C \frac{i_s}{G} = jQ \times i_s \tag{4.27}$$

where we have defined the circuit Q factor at resonance by

$$Q = \frac{\omega_0 C}{G} \tag{4.28}$$

in complete analogy with Eq. 4.3. Likewise, the current gain through the inductor is also easily derived

$$i_L = -jQ \times i_s \tag{4.29}$$

The equivalent expressions for the circuit Q factor are given by the inverse of the relations of Eq. 4.5

$$Q = \frac{\omega_0 C}{G} = \frac{R}{\omega_0 L} = \frac{R}{\frac{1}{\sqrt{LC}} L} = \frac{R}{\sqrt{\frac{L}{C}}} = \frac{R}{Z_0} \tag{4.30}$$

The phase response of a resonant circuit is also related to the Q factor. For the parallel RLC circuit the phase of the admittance is given by

$$\angle Y(j\omega) = \tan^{-1}\left(\frac{\omega C\left(1 - \frac{1}{\omega^2 LC}\right)}{G}\right) \tag{4.31}$$

The rate of change of phase at resonance is given by

$$\left.\frac{d\angle Y(j\omega)}{d\omega}\right|_{\omega_0} = \frac{2Q}{\omega_0} \tag{4.32}$$

A plot of the admittance phase as a function of frequency and Q is shown in Fig. 4.6.

Circuit Transfer Function

Given the duality of the series and parallel RLC circuits, it's easy to deduce the behavior of the circuit. Whereas the series RLC circuit acted as a filter and was only sensitive to voltages near resonance ω_0, likewise the parallel RLC circuit is only sensitive to currents near resonance

$$H(j\omega) = \frac{i_o}{i_s} = \frac{v_o G}{v_o Y(j\omega)} = \frac{G}{j\omega C + \frac{1}{j\omega L} + G} \tag{4.33}$$

which can be put into the same canonical form as before

$$H(j\omega) = \frac{j\omega \frac{\omega_0}{Q}}{\omega_0^2 + (j\omega)^2 + j\frac{\omega\omega_0}{Q}} \tag{4.34}$$

where we have appropriately re-defined the circuit Q to correspond the parallel RLC circuit. Notice that the impedance of the circuit takes on the same form

$$Z(j\omega) = \frac{1}{Y(j\omega)} = \frac{1}{j\omega C + \frac{1}{j\omega L} + G} \tag{4.35}$$

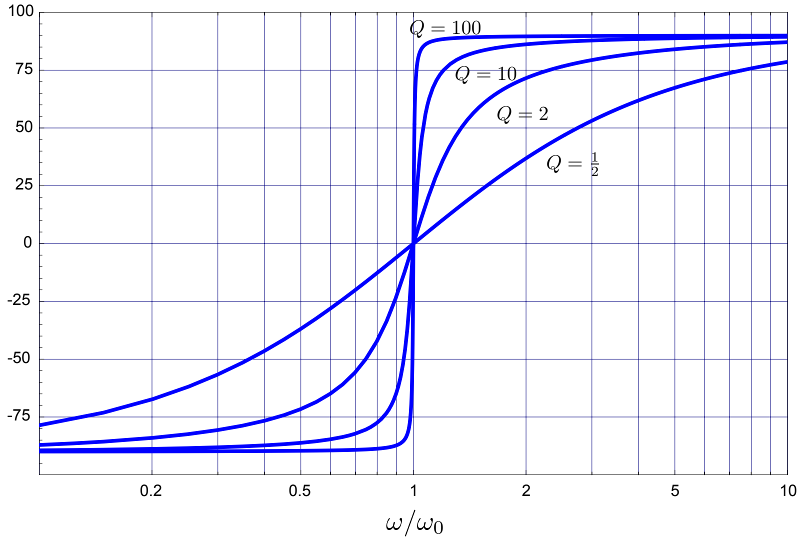


Figure 4.6: The phase of a second order admittance as function of frequency. The rate of change of phase at resonance is proportional to the *Q* factor.

which can be simplified to

$$Z(j\omega) = \frac{j\frac{\omega}{\omega_0}\frac{1}{GQ}}{1+\left(\frac{j\omega}{\omega_0}\right)^2 + j\frac{\omega}{\omega_0 Q}} \tag{4.36}$$

At resonance, the real terms in the denominator cancel

$$Z(j\omega_0) = \frac{j\frac{R}{Q}}{\underbrace{1+\left(\frac{j\omega_0}{\omega_0}\right)^2}_{=0} + j\frac{1}{Q}} = R \tag{4.37}$$

It's not hard to see that this circuit has the same half power bandwidth as the series *RLC* circuit, since the denominator has the same functional form

$$\frac{\Delta\omega}{\omega_0} = \frac{1}{Q} \tag{4.38}$$

A plot of this impedance versus frequency has the same form as Fig. 4.4 multiplied by the resistance *R*.

Energy storage in a parallel *RLC* circuit is completely analogous to the series *RLC* case and in fact the general equation relating circuit *Q* to energy storage and dissipation also holds in the parallel *RLC* circuit.

4.2.5 The Many Faces of Q

As we have seen, in *RLC* circuits the most important parameter is the circuit *Q* and resonance frequency ω_0. Not only do these parameters describe the circuit in a general way, but they also give us immediate insight into the circuit behavior.

The *Q* factor can be computed several ways, depending on the application. For instance, if the circuit is designed as a filter, then the most important *Q* relation is the half-power bandwidth

$$Q = \frac{\omega_0}{\Delta\omega} \tag{4.39}$$

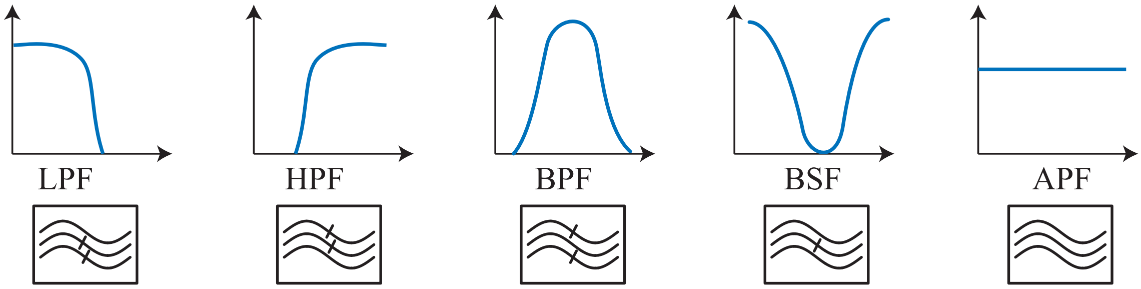


Figure 4.7: Standard filter type include the (a) low-pass filter, (b) high-pass filter, (c) band-pass filter, (d) band reject filter, and (e) all-pass filter.

We shall also find many applications where the phase selectivity of these circuits is of importance. An example is a resonant oscillator where the noise of the system is rejected by the tank based on the phase selectivity. In an oscillator any "excess phase" in the loop tends to move the oscillator away from the natural resonant frequency. It is therefore desirable to maximize the rate of change of phase of the circuit impedance as a function of frequency. For the parallel *RLC* circuit we derived the phase of the admittance (Eq. 4.32) which gives us another way to interpret and compute Q

$$Q = \frac{\omega_0}{2} \frac{d\angle Y(j\omega)}{d\omega} \tag{4.40}$$

For applications where the circuit is used as a voltage or current multiplier, the ratio of reactive voltage (current) to real voltage (current) is most relevant. For a series case we found

$$Q = \frac{v_L}{v_R} = \frac{v_C}{v_R} \tag{4.41}$$

and for the parallel case

$$Q = \frac{i_L}{i_R} = \frac{i_C}{i_R} \tag{4.42}$$

The last and one of the most important interpretations of Q is in the definition of energy, relating the energy storage and losses in a *RLC* "tank" circuit. We can define the of Q a circuit at frequency ω as the energy stored in the tank W divided by the rate of energy loss

$$Q = W / \frac{dW}{d\phi} = \omega W / \frac{dW}{dt} \tag{4.43}$$

4.3 Filter Frequency Response

Filters are key building blocks in communication systems. Common filters that you are no doubt familiar with include low-pass filters (LPF), high-pass filters (HPF), and band-pass filters (BPF). Less common, but equally important, include band-reject filters and all-pass filters. Common notation and ideal filter magnitude responses for these various filters are shown in Fig. 4.7. The all-pass filter may seem useless at first, but is actually quite useful when we examine the filter's phase transfer characteristic. The filter can be used to equalize the phase response of a distorted signal.

A LPF passes the lower frequencies to the output and attenuates the high frequency components beyond the cut-off frequency. The simplest low-pass filter consist of an *RC* circuit shown in Fig. 4.8a, which attenuates the signal at 20 dB/decade beyond the cutoff, a rather gentle roll-off. To

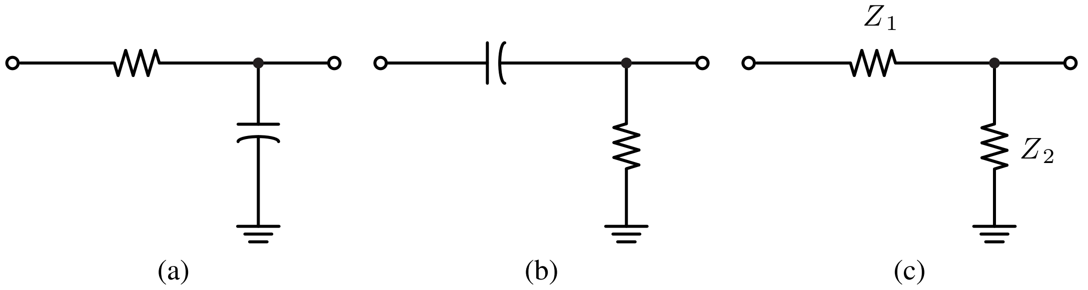


Figure 4.8: (a) A simple *RC* low-pass filter. (b) A simple *RC* high-pass filter. (c) An arbitrary filter viewed as a voltage divider.

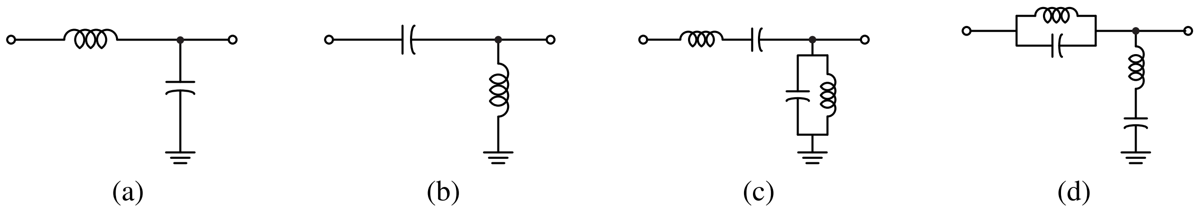


Figure 4.9: (a) An *LC* low-pass filter. (b) An *LC* high-pass filter. (c) An *LC* pass band filter. (d) An *LC* band reject filter.

improve the roll-off, the *LC* filter shown in Fig. 4.9a can be used. Intuitively, at low frequencies the inductor is a short and the capacitor is open, so the signal is coupled to the output. At resonance the voltage across the inductor/capacitor equal, which sets the 3-dB frequency. At high frequencies, though, the inductor reactance increases and the capacitor reactance decreases, decoupling the input from the output and tending to short the output signals to ground. The transfer function of this filter is given by analyzing the circuit as a voltage divider (Fig. 4.8c), where $Z_1 = Z_C||R_L$ and $Z_2 = j\omega L + R_S$

$$\frac{V_2}{V_1} = \frac{Z_C||R_L}{Z_C||R_L + j\omega L + R_S} \tag{4.44}$$

Assume that $R_0 = R_L = R_S$

$$= \frac{1}{1+(1+j\omega L/R_0)(1+j\omega R_0 C)} \tag{4.45}$$

which shows that the filter has two complex poles, causing a roll-off of 40 dB/dec. The steepness and bandwidth of the filter near the cutoff frequency is controlled quality factor Q of the transfer function.

This filter can be converted to a high-pass filter by interchanging the inductor and the capacitor, which rejects the low frequencies due to the capacitor coupling and the inductor shunting the signal to ground, and passes the high frequencies, as shown in Fig. 4.9b. How do we realize a band-pass filter? If we view the circuit as an arbitrary voltage divider, notice that in the passband of the filter the impedance Z_2 is shorted and the impedance Z_1 is opened whereas the opposite occurs ideally in the stop-band. A short circuit is realized at an arbitrary frequency using a series *LC* circuit and an open circuit is realized with a parallel *LC* circuit. Putting these ideas together leads to Fig. 4.9c, a passband filter.

By the same argument, a band reject filter is realized by interchanging the role of the Z_1 and Z_2, as shown in Fig. 4.9d. We see that simple *LC* resonant circuits are extremely versatile and form

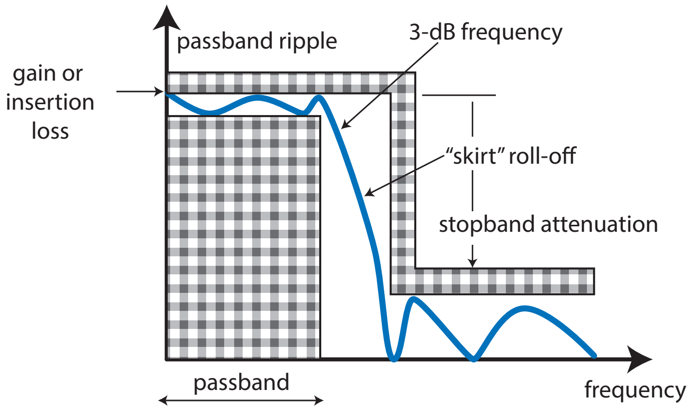


Figure 4.10: The desired filter response "mask" and the transfer characteristics of a filter.

the core of an entire family of filters. As we shall see this powerful insight can be extended even further.

How do we select the various component values to realize a given filter response? Well, the answer depends on what you are trying to achieve. A filter is typically characterized by specifying the filter mask shown in Fig. 4.10. The mask is characterized by the following parameters: corner frequency or 3-dB bandwidth, pass-band ripple, which measures how much the in-band signal gain varies (which leads to distortion), the stop-band rejection, and the steepness of the "skirt" of the filter (the transition region of the filter). The filter roll-off is related to the filter order, or the number of poles. Other important metrics include the group delay (see below), insertion loss of the filter (rather than the ripple), and the stop-band ripple (only for certain filter types). Some filters have transmission zeros, which cause the filter response to go to zero at specific frequencies in the stop-band. Insertion loss is related to the fact that real inductors/capacitors have loss, and some of the input energy is absorbed by the network and converted to heat.

An ideal filter would delay all frequency components of the signal by the same amount, i.e. the filter would have a constant *group delay*

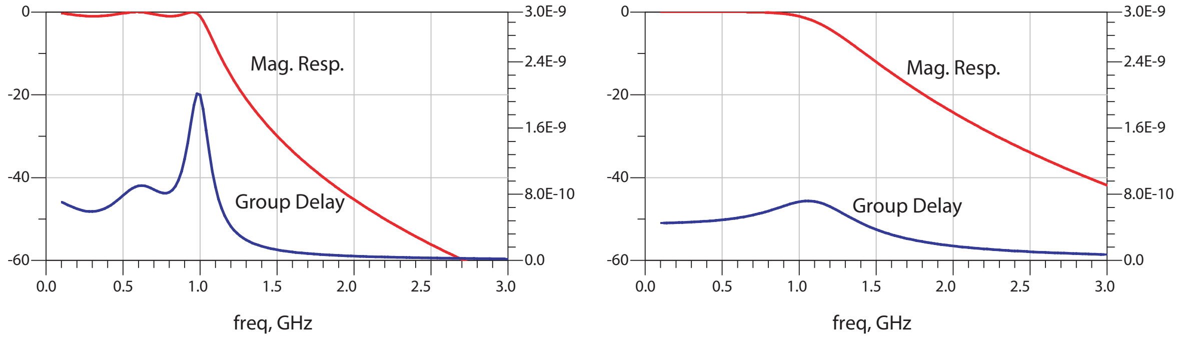
$$\tau_g = -\frac{d\angle H}{d\omega} \tag{4.46}$$

Any variation in the group delay leads to distortion since different components of the signal arrive at different times. Notice that an ideal filter should therefore have a linear phase response or a flat (constant) group delay. To see this, notice that a distortionless filter should preserve the waveform shape, which means that the output can only be a scaled and delayed version of the input signal (for the band of interest). The transfer function for such a filter takes the form

$$H(s) = |H_0|e^{-sT} \tag{4.47}$$

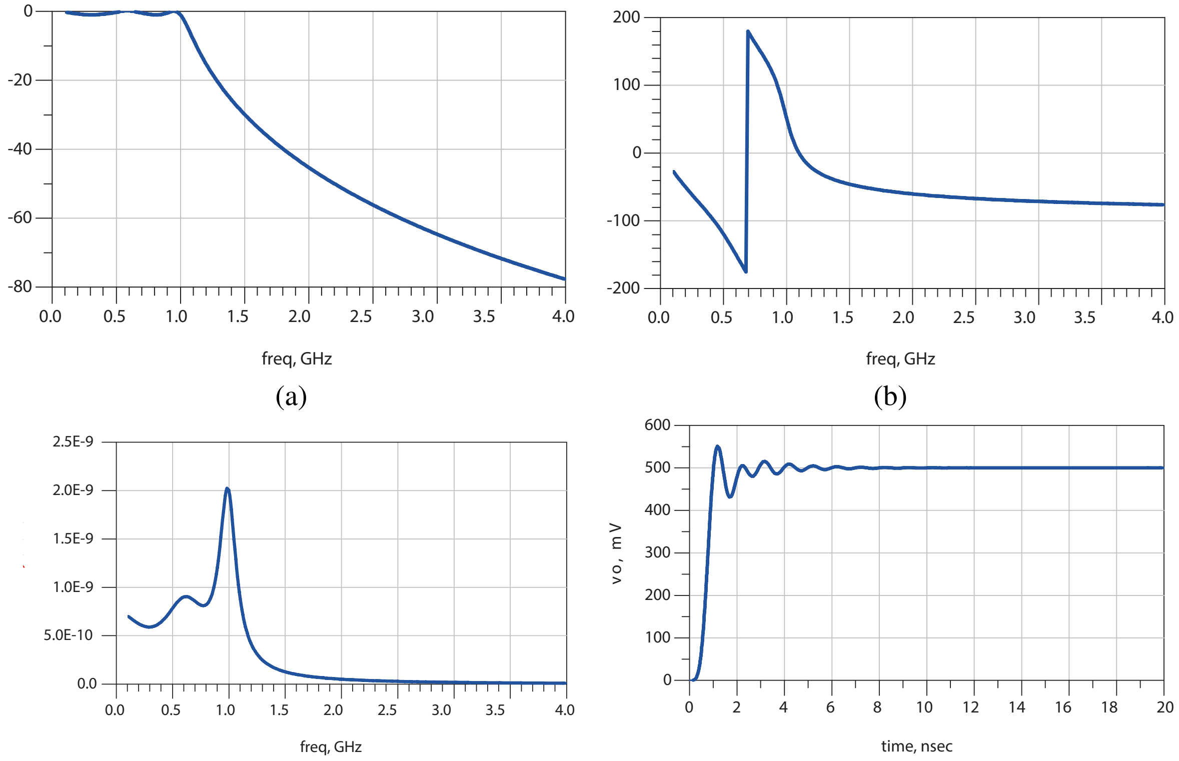
where H_0 is the scaling factor and T is the delay of the filter. It is therefore important for the filter approximate this response in the band of interest, which means minimizing the ripple in the amplitude response and realizing a linear phase response (or constant group delay).

In general there is a trade-off in the filter attenuation characteristics and the group delay, which means that more out-of-band magnitude attenuation results in more phase distortion in-band and vice-versa. In Fig. 4.11 we compare two filters which differ in their attenuation rate; notice that the filter with higher attenuation has considerably worse group delay variation. An all-pass filter can be used to compensate for the phase distortion of a given filter, or to compensate for the phase distortion of a transmission medium (such as a cable or device with poor frequency response).



(a) (b)

Figure 4.11: (a) A sharp roll-off filter has more than 40-dB attenuation at 2 GHz and poor group delay (Chebyshev order 5) whereas (b) a soft roll-off filter has much smaller group delay (Butterworth order 5).



(a) (b)

(c) (d)

Figure 4.12: (a) The magnitude, (b) phase response, (c) group delay and (d) step response of a filter.

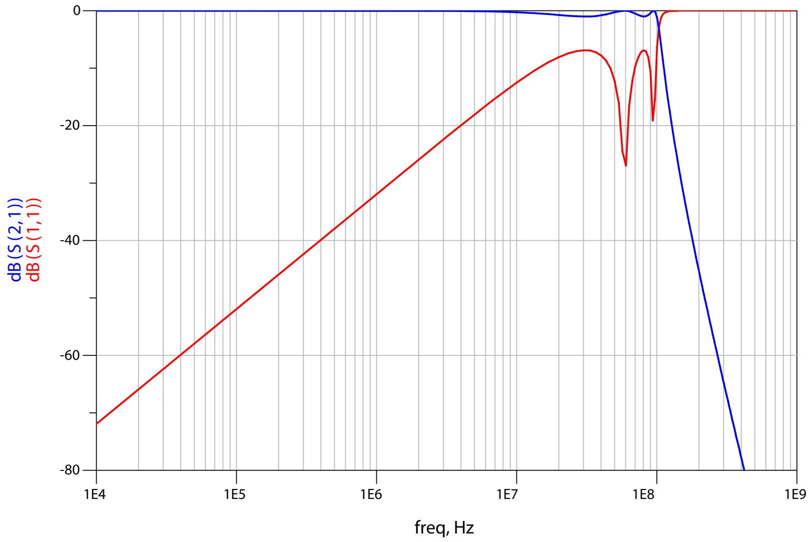


Figure 4.13: The input reflection coefficient s_{11} and transfer coefficient s_{21} for a fifth-order Chebyshev filter.

A simple way to observe the distortion caused by the non-constant group delay is to plot the step response of the filter. Since the step transition has high frequency components which must all arrive at the same instant, any deviation from a linear phase response leads to distortion in the waveform, as shown in Fig. 4.12.

Filters are two-port elements and thus a full characterization requires the specification of four complex parameters. If a filter is realized with only passive elements, then the two-port is reciprocal and $z_{12} = z_{21}$. Many filters are also symmetric so $z_{11} = z_{22}$. In these cases the filter is fully characterized by two complex frequency dependent parameters. Most commonly filters are characterized in terms of their scattering parameters since this is how the filters are measured. The input reflection coefficient, or s_{11}, is particulary important since the energy transferred into the filter is given by

$$P_{in} = 1 - |\Gamma(\omega)|^2 = 1 - |s_{11}|^2 \tag{4.48}$$

which means in the pass-band we desire $|s_{11}| \approx 0$ (no power reflected) whereas in the stop-band we desire $|s_{11}| \approx 1$, which means that all the incident power is reflected. If the filter is realized with lossless or low-loss elements, then the input power is actually the power delivered to the load. A filter characterized in this way is shown in Fig. 4.13, where the input reflection s_{11} and transmission s_{21} are plotted versus frequency.

The filter transfer characteristics are measured or calculated through s_{21}. For an ideal filter $|s_{21}| \approx 1$ in the passband and $|s_{21}| \approx 0$ in the stop-band. For any real filter, there is some insertion loss due to the inevitable resistance in the components, and the magnitude of $|s_{21}|$ in the passband indicates this loss (under matched conditions).

4.3.1 Ladder Filters

The concept of a voltage divider can be extended, as shown in Fig. 4.14a, to even realize higher out-of-band attenuation. The buffer is used to isolate the two filters and so the overall transfer function is a cascade of the individual transfer functions, doubling the attenuation from 40 dB/dec to 80 dB/dec, a significant improvement. In actual practice, the same effect can be realized without the buffer, as shown in Fig. 4.14b, except now the transfer function is more complicated but has the

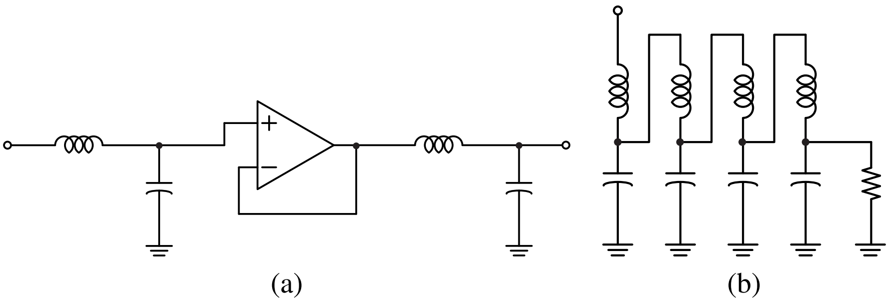


Figure 4.14: (a) Cascading simple filters results in higher order roll-off characteristics. (b) Cascaded filters without the buffer.

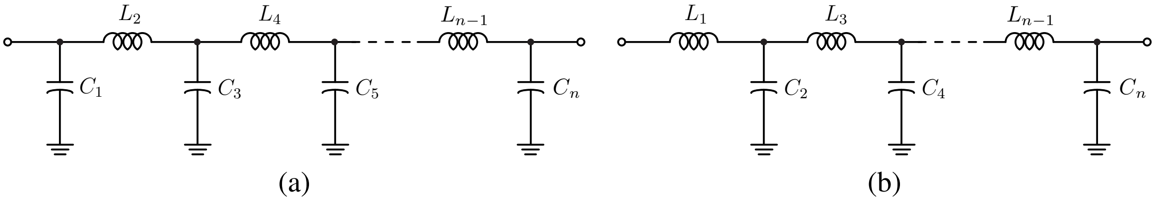


Figure 4.15: (a) The canonical ladder low-pass filter of order *n* (odd). (b) The canonical ladder filter of order *n* (even).

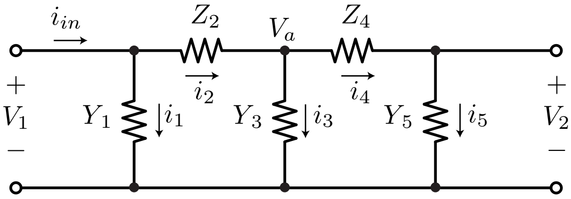


Figure 4.16: (a) An arbitrary fifth-order ladder filter structure.

same order. All of the filters discussed up to this point can in fact be extended in this fashion to realize higher order filters. The order of the filter correspond to the number of "rungs" in the ladder filter, redrawn in standard form in Fig. 4.15. This canonical ladder filter structure can be converted from low-pass to high-pass, band-pass, or band-stop by the following simple transformations:

- LP → HP: $L \to C, C \to L$
- LP → BP: $L \to$ series $LC, C \to$ parallel LC
- LP → BS: $L \to$ parallel $LC, C \to$ series LC

The two-port parameters of an arbitrary ladder filter shown in Fig. 4.16 can be calculated by noting that the input impedance is given by[1]

$$Z_{11} = \cfrac{1}{y_1 + \cfrac{1}{z_2 + \cfrac{1}{y_3 + \cfrac{1}{z_4 + \cdots}}}} \tag{4.49}$$

To calculate the transfer characteristic Z_{21}, leave the output open-circuited and assume the voltage $V_2 = 1\text{V}$ and find the input current step by step [**Aatre**]. Note that $i_5 = y_5 V_2 = y_5$ and $i_4 = i_5$ so we have

$$V_a = i_4 z_4 + V_2 = 1 + z_4 y_5 \tag{4.50}$$

Repeating this calculation and using $i_2 = i_3 + i_4$

$$i_3 = y_3 V_a = y_3 + y_3 z_4 y_5 \tag{4.51}$$

$$V_1 = (i_3 + i_4) z_2 + V_a = z_2 y_3 + z_2 y_3 z_4 y_5 + z_2 y_5 + z_4 y_5 + 1 \tag{4.52}$$

The input current is given by $I_1 = y_1 V_1 + i_2$, which leads to

$$y_{21} = y_1 z_2 y_3 + y_1 z_2 y_3 z_4 y_5 + y_1 z_2 y_5 + y_1 z_4 y_5 + y_1 + y_3 + y_3 z_4 y_5 + y_5 \tag{4.53}$$

In practice carrying out the algebra is unnecessary since filters have been studied extensively and canonical filter structures have been tabulated and are widely available [**matthie**]. Filter design tools are also abundant on the web and through specialized software packages (such as ADS). Nevertheless you are encouraged to play around with a few simple filters to gain intuition before using the tools.

Impedance Matching

It is now clear that in a standard band-pass filter structure, all the LC component values are chosen to resonate at the center frequency of the filter response, which constrains the product of LC. The ratio of these components must be chosen carefully to produce the desired filter properties. For instance, in order to obtain an impedance match, the input impedance of the filter should equal the desired load impedance across the passband. If we view a simple three section low-pass matching network as to front-to-back "L" matching networks (bisecting the series element as shown in Fig. 4.17), then we view this as a classic impedance down transformation by the factor $(1+Q_2^2)^{-1}$ followed by an up transformation of $(1+Q_1^2)$ resulting in an overall transformation of (at resonance)

$$Z_{in} = \frac{1+Q_1^2}{1+Q_2^2} \cdot R_L \tag{4.54}$$

[1] In fact, this form if very useful for synthesis of a ladder filter of a given transfer function. Using long division, a transfer function can be written in continued fraction form, and the element values are readily calculated.

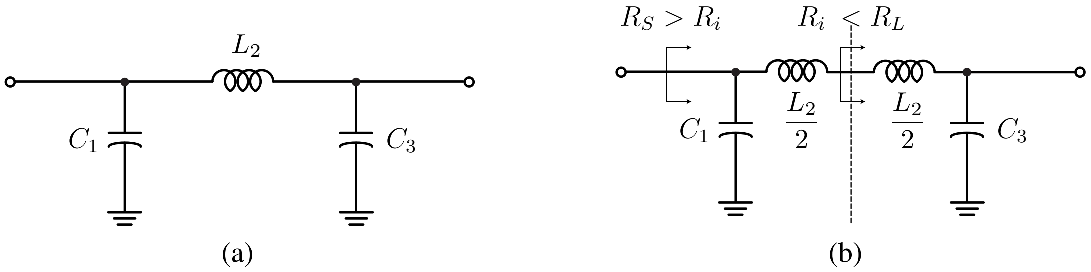


Figure 4.17: (a) A three section low-pass filter. (b) The series element is bisected to form two “L” matching networks.

Which shows that the filter can amplify or attenuate the voltage, or in other words change the impedance seen by the source. This impedance matching property of the filter is useful if the source and load impedance are different. In many applications, though, these are the same so ideally $Z_{in} = R_L$ at resonance, which means we should choose the filter component values to satisfy this constraint, or $Q_1 = Q_2$. The overall transfer characteristics of the filter come down to one free parameter, the quality factor Q.

4.3.2 Standard Filter Families

Filters have been thoroughly analyzed and classified into families of filters with specific filter characteristics. These various filters trade-off passband ripple for sharper attenuation or slightly worse group delay. The terminology behind these filters is widely known and standardized (although the spelling *Cheybchev* varies) and the names of the filters derive from the original mathematicians who studied the functions that underlie these filter transfer characteristics.

Butterworth Filters

The simplest of the family of filters, the Butterworth filters are also known as “maximally flat”, since the transfer function

$$\left|\frac{V_2}{V_1}\right| = \frac{1}{\sqrt{1+(f/f_0)^{2n}}} \tag{4.55}$$

has n zero derivatives at the origin. This means that the filter response remains as flat as possible in the passband. As evident in Fig. 4.18, the order of the filter n determines the stop-band rolloff. To realize such a filter requires exactly n elements in the ladder structure. It is relatively easy to show that the poles of this filter lie uniformly on a half-circle in the left-hand plane with radius ω_0.

Chebyshev Filters

The Chebyshev filter has a sharper roll-off compared to the Butterworth filter in the stop-band, as shown in Fig. 4.19. The trade-off is that the Chebyshev filter introduces ripple in the passband. The transfer function for this filter is given by

$$\left|\frac{V_2}{V_1}\right| = \frac{1}{\sqrt{1+\varepsilon^2 T_n^2(f/f_0)}} \tag{4.56}$$

where $T_n(x)$ is a Chebyshev polynomial of order n. This filter is realized with n elements. The in-band ripple is controlled by adjusting the factor ε. For a given value of ε, the in-band ripple is given by

$$\text{ripple in dB} = 20\log_{10}\frac{1}{1+\varepsilon^2} \tag{4.57}$$

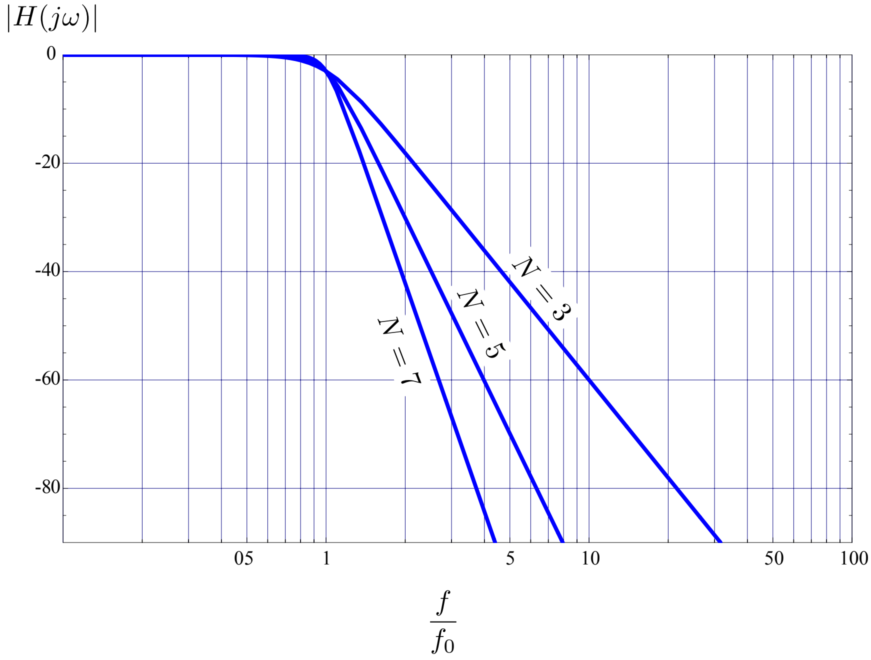


Figure 4.18: The voltage transfer characteristics of a Butterworth filter. The roll-off slope varies with the filter order n.

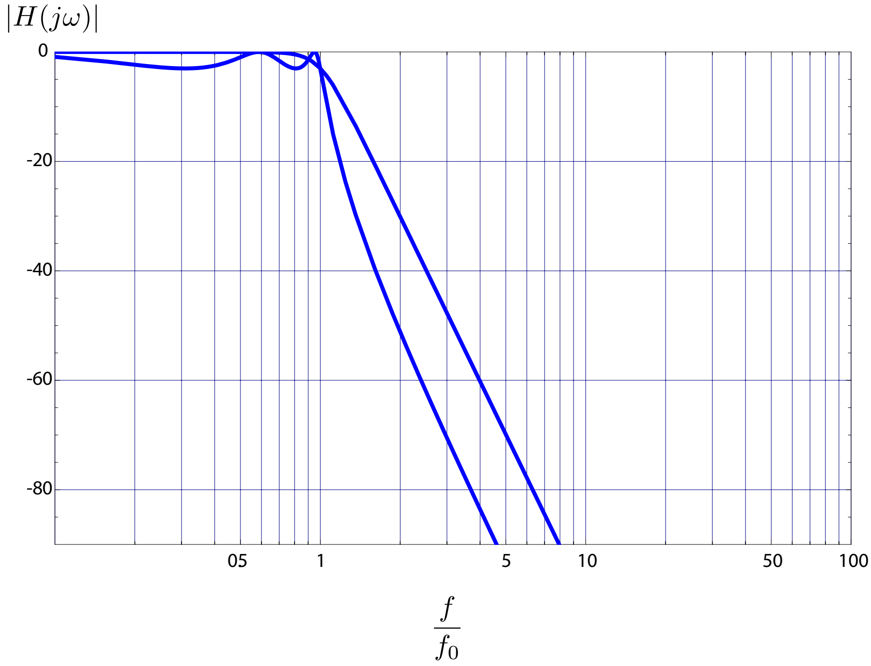


Figure 4.19: The voltage transfer characteristics of a Chebyshev and Butterworth filter. The Chebyshev has faster roll-off and in-band ripple whereas the Butterworth filter has a flat in-band response.

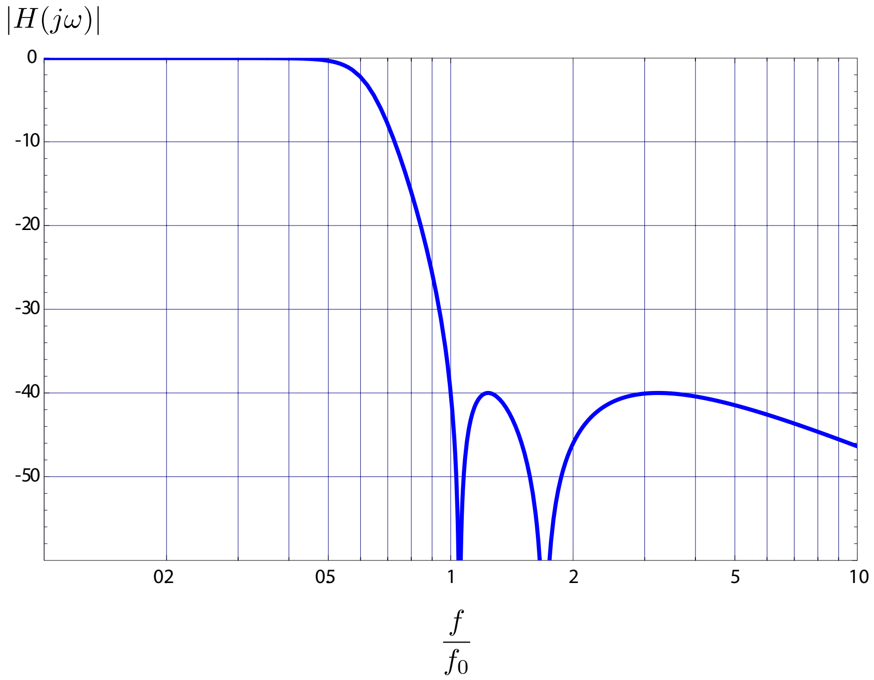


Figure 4.20: The voltage transfer characteristics of a type-II or Inverse Chebyshev filter. The Inverse Chebyshev has good out-of-band rejection.

The Chebyshev polynomial is given by

$$T_n(x) = \cosh(n \operatorname{arccosh}(x)) \tag{4.58}$$

It is not too difficult to show that the poles of this filter are distributed on the left-hand side of the s-plane along an ellipse[**wikicheby**].

Other Filter Families

In the literature you will encounter other filter families that display various other trade-offs. For instance, the Bessel filter has a maximally flat group delay, which is ideal for applications intolerant to phase distortion in the passband. Inverse Chebyshev filters have no ripple in the passband but have ripple in the stop band instead, as shown in Fig. 4.20. These filters have both poles and zeros in their transfer characteristics. Elliptical filters allow one to specify passband and stopband ripple and achieve very good stop-band attenuation, but require high-Q poles. For all filter families it's important to remember that they are all realized by using the ladder filter structure. Only the components values vary to change the transfer function from one filter type to another.

4.3.3 Filter Transformations

Most filters families are tabulated as low-pass filters and normalized for a cutoff frequency of $\omega_c = 1$ radian/sec. Two types of filters can be used, one beginning with a shunt capacitor or series inductor (Fig. 4.15). The component values are given as g_n (Farads/Henrys), assuming a source impedance of $R_s = 1\Omega$ and load impedance $R_L = 1\Omega$ (odd-order filters). For even-order filters, $R_L = 1/g_{n+1}$.

Frequency and impedance scaling

The low-pass filter cutoff frequency can be scaled to ω_c and scaled to a source impedance of R_0 by modifying g_n in the following way

$$L_n = \frac{R_0 g_n}{\omega_c} \tag{4.59}$$

$$C_n = \frac{g_n}{R_0 \omega_c} \tag{4.60}$$

Low-pass to high-pass transformation

The frequency substitution $-\omega_c/\omega \rightarrow \omega'$ converts the filter prototype from low-pass to high-pass. The new component values are given by

$$C'_n = \frac{1}{g_n} \tag{4.61}$$

$$L'_n = \frac{1}{g_n} \tag{4.62}$$

Band-pass and band-stop transformation

The frequency substitution $\frac{\omega_0}{\omega_2-\omega_1}\left(\frac{\omega}{\omega_0}-\frac{\omega_0}{\omega}\right) \rightarrow \omega'$ is used to tranform the low-pass filter to a bandpass filter. The fractional bandwidth Δ of the filter is given by

$$\Delta = \frac{\omega_2 - \omega_1}{\omega_0} \tag{4.63}$$

The center frequency is the geometric (not arithmetic) mean of the 3-dB frequencies

$$\omega_0 = \sqrt{\omega_1 \omega_2} \tag{4.64}$$

Carrying out the arithmetic means that a series inductor is transformed into a series *LC* circuit with component values

$$L'_n = \frac{L_n R_0}{\omega_0 \Delta} \tag{4.65}$$

$$C'_n = \frac{\Delta}{\omega_0 L_n R_0} \tag{4.66}$$

and the shunt capacitors are transformed into a parallel *LC* circuit with components given by

$$L'_n = \frac{R_0 \Delta}{\omega_0 C_n} \tag{4.67}$$

$$C'_n = \frac{C_n}{\omega_0 R_0 \Delta} \tag{4.68}$$

The inverse transformation $\Delta\left(\frac{\omega}{\omega_0}-\frac{\omega_0}{\omega}\right)^{-1} \rightarrow \omega'$ is used to realize a bandstop filter. The series inductors are converted to parallel *LC* branches

$$L'_n = \frac{R_0 L_n \Delta}{\omega_0} \tag{4.69}$$

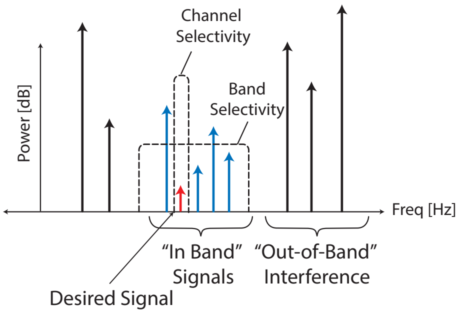


Figure 4.21: A radio receiver must be able to discriminate between many undesired and large interfering signals in order to "hear" a weak desired signal. Each arrow respresents a modulated signal at a given frequency.

$$C_n' = \frac{1}{\omega_0 L_n R_0 \Delta} \tag{4.70}$$

whereas the shunt capacitors are converted into series *LC* branches

$$L_n' = \frac{R_0}{\omega_0 C_n \Delta} \tag{4.71}$$

$$C_n' = \frac{C_n \Delta}{\omega_0 R_0} \tag{4.72}$$

4.3.4 Filters in Communication Systems

In a wireless communication system filters are used to "pick a needle out of a haystack," or in other words to select the desired signal of interest in a sea of interfering signals. This is shown schematically in Fig. 4.21, where the desired signal is at channel 2, or 910 MHz, in a radio band that can support many different channels. The first filter in this example is a fixed filter that selects all the channels of interest while rejecting "out of band" signals, such as the multitude of in other frequency bands, such as UHF television up to 600 MHz and other cellular and wireless communication signals in the 2-5 GHz spectrum. The second filter is used to further isolate the desired channel from the rest of the signals to provide channel selectivity. A difficult, but typical situation for a receiver is the "near-far" problem, shown in Fig. 4.22, where a multitude of interferers appear around the desired signal which is weak. If the receiver does not have sufficient sensitivity and linearity, these interfering signals can jam a receiver.

Notice that the second narrow filter needs to have a variable center frequency unless we use static channel assignments, which is very unrealistic in practice. Since in practice it is much easier to build a high quality fixed center frequency filter rather than a tunable filter, the most popular way to realize the same effect is to downconvert the signal of interest to a fixed intermediate frequency (IF), where channel selection and blocker attenuation can be performed. This is the basis of the superheterodyne receiver architecture, shown in Fig. 4.23. Notice that there are three filters used in this architecture, an RF band select filter, an image-reject filter, the importance of which will be highlighted later, and the a second IF filter. In this lab you will design these filters, simulate them, and then build and test the filters.

In a high-speed communication system using amplitude modulation, the time-domain ripple at the amplitude transitions leads to inter-symbol interference (ISI) and "eye closure". In Fig. 4.24 we

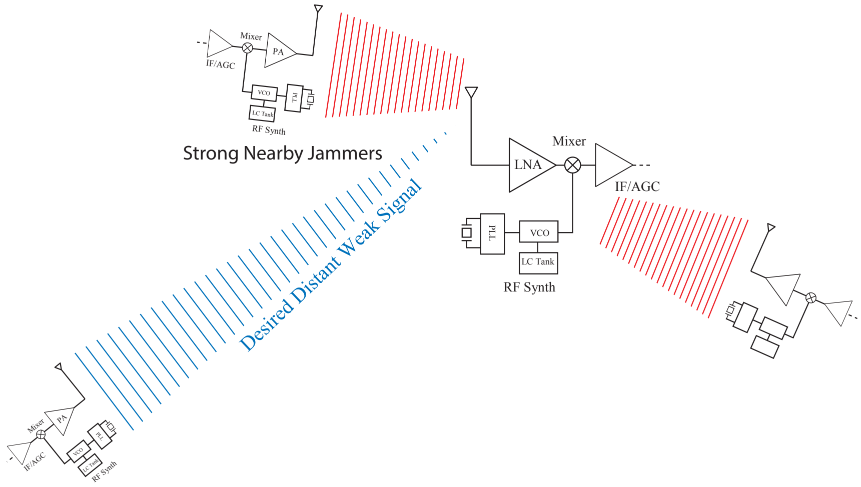


Figure 4.22: A radio receiver must contend with a near-far scenario where the desired transmitter is far away resulting in a weak received signal in the presence of many nearby unwanted interferers.

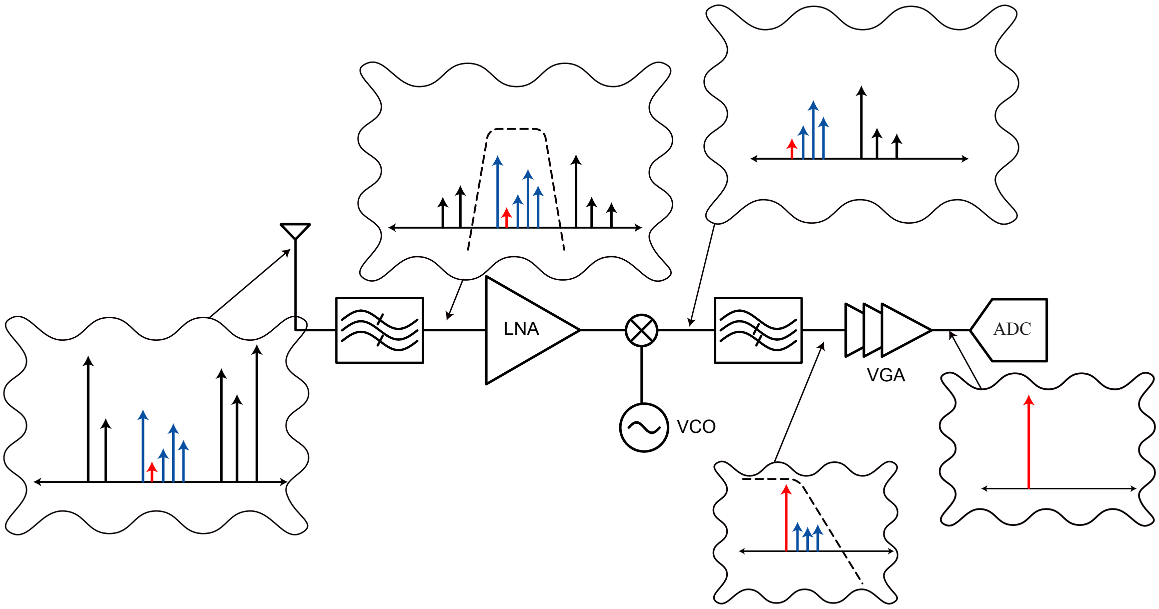


Figure 4.23: A superheterodyne radio receiver architecture uses several band-pass, band-reject, and low-pass filters.

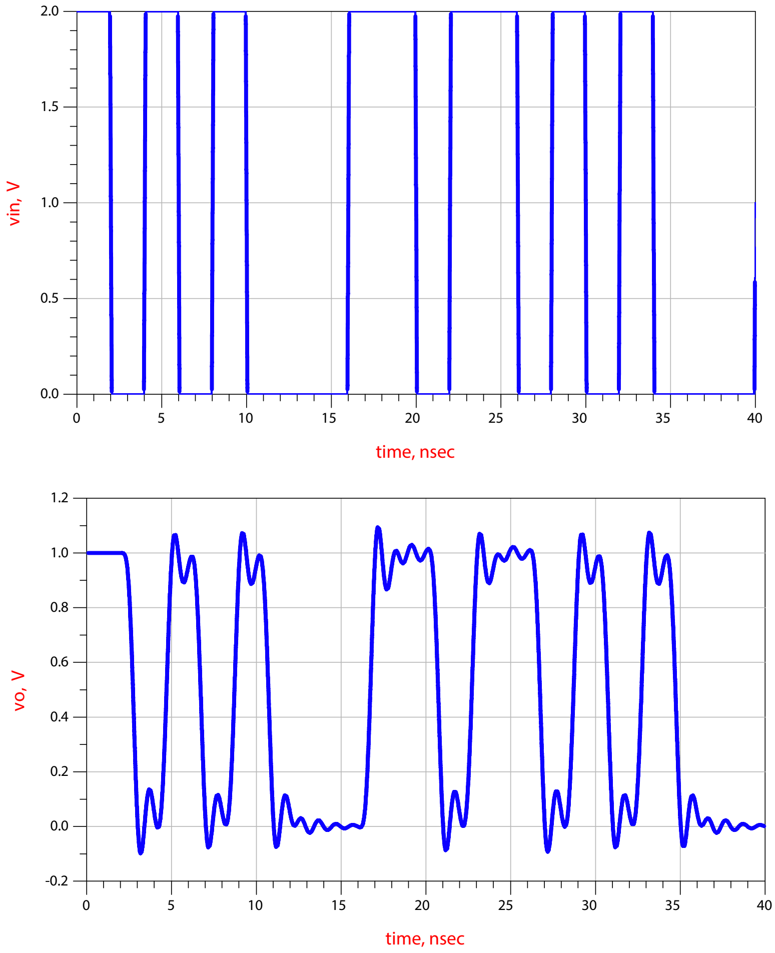


Figure 4.24: Input pulse data waveform (top) and the filtered response (bottom) shows significant inter-symbol interference (ISI).

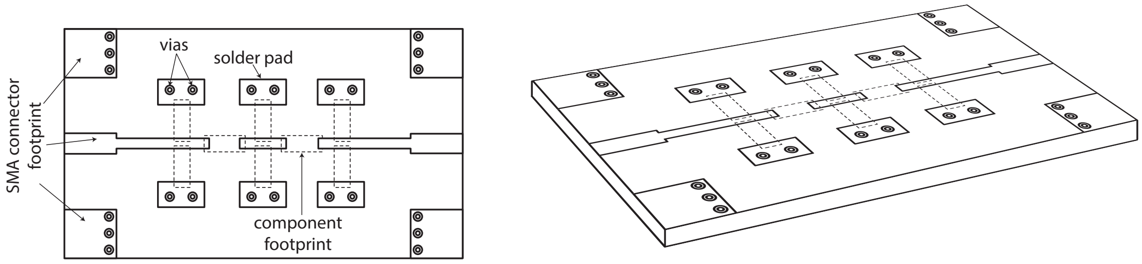


Figure 4.25: The printed circuit board layout for a ladder filter structure.

compare the pulse response of a Bessel and Chebyshev filter. Note that in both cases we lose the sharp edges due to filtering, but the Bessel filter does not ring, which prevents "leakage" of one bit onto another.

4.4 Appendix

In this appendix we discuss how you can build filters using planar PCB manufacturing process and lumped surface mount components.

4.4.1 PCB Manufacturing

A printed circuit board (PCB) consists of a stack of dielectric materials, usually FR4 ($\epsilon_r = 4.4$), with a typical thickness on the order of 50 mils. The thicknesses can be varied either for more mechanical stability or even thinned to obtain flexibility. Each dielectric is sandwiched by two layers of Cu metal, which can be used as interconnect or as a ground plane. In practice, two or more layers are used to build the PCB. Even if many layers are used, one layer is reserved for RF signals, with a nearly solid ground plane serving as the RF return path and a single layer forming RF transmission lines. The metal layers have a typical thickness of 30 μm. In a single layer PCB, the backside of the board is a solid ground plane. Connections to ground must travel through a "via" to reach the backside.

Consider the layout of the board shown in Fig. 4.25. The input and output of the board have footprints for SMA connectors which allow you to connect SMA cables. The input and output microstrip transmission lines are interrupted periodically which allows you to place components in series or in shunt in a ladder filter structure. Landing pads with vias to ground also appear periodically to allow shunt components to be soldered to ground.

To solder components onto the board, you use standard surface mount components (SMT), known as SMT elements. SMT components are usually classified according to their footprint size in mils, for example 0603 means 6 mils by 3 mils. To solder these components by hand, you must use a sharp solder needle and a microscope to do the soldering. If you have an advanced soldering iron with temperature control, use the manufacture recommend setting.

4.4.2 Lumped Components

Up to now you have designed your filter with ideal components (inductors, capacitors, resistors), but real circuit components are far from ideal. Due to their physical size, no real component is truly "lumped", so every device is marred by parasitics. Consider, for instance, a capacitor, which has an equivalent circuit model shown in Fig. 4.26. The model has many parasitic components which only become relevant at high frequencies. A plot of the impedance of the capacitor, shown in Fig. 4.27,

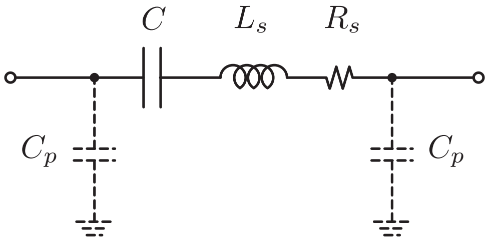


Figure 4.26: The lumped equivalent circuit model for a real soldered capacitor.

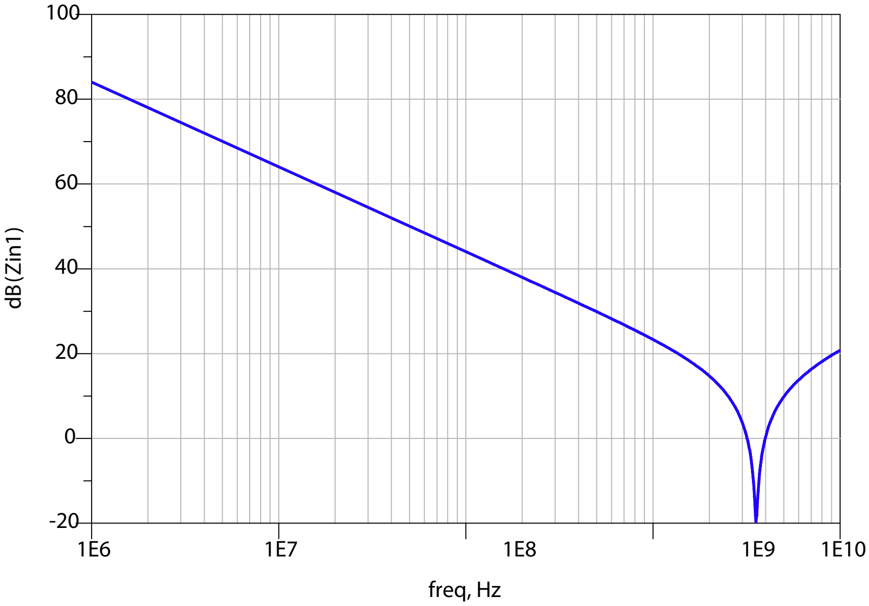


Figure 4.27: The magnitude of the impedance of a real capacitor.

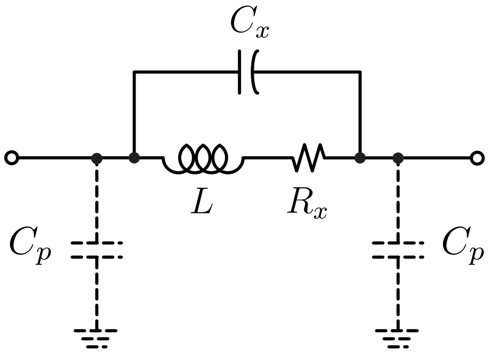
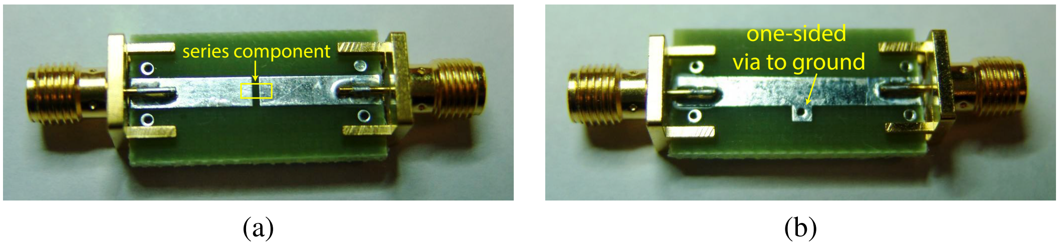


Figure 4.28: The lumped equivalent circuit model for a real soldered inductor.



(a)

(b)

Figure 4.29: (a) Open and (b) short structures used for the automatic port extension feature of the network analyzer.

shows that in addition to the ideal behavior, the most notable difference is the self-resonance that occurs for any real capacitor. The self-resonance is inevitable for any capacitor of finite dimension, since the current that flows into the plates of the capacitor stores magnetic energy, and at a high enough frequency, this energy is equal to the electric energy stored by the capacitor, resulting in resonance. This inductance is exacerbated by the leads of the capacitor, which can often dominate the inductance. The inductive parasitics are lumped into a single inductor L_s in series with the capacitor. The finite conductivity of the plates and the leads also results in some series loss, modeled by R_s (sometimes labeled *ESR*, or effective series resistance). Unless a capacitor is fabricated in a vacuum, the dielectric material that separates the plates also has loss (and resonance), which is usually modeled by a large shunt resistance R_{die}. Furthermore, when a capacitor is soldered onto a PCB, there is parasitic capacitance from the solder pads to the ground plane, resulting in the capacitors C_p in the equivalent model.

In a like manner, every inductor also has parasitics, as shown in the equivalent circuit model (Fig. 4.28), which limit its operating frequency range. The series resistance R_x is due to the winding resistance and the capacitance C_x models the distributed turn-to-turn capacitance of the structure in a lumped element. The element self resonates at a frequency of approximately $1/\sqrt{LC_x}$ and has a quality factor $Q = \omega L/R_x$. When the inductor is soldered onto the PCB, there is an additional capacitance to ground modeled by C_p, which lowers the self-resonant frequency to $1/\sqrt{L(C_x + C_p/2)}$.

In addition to the component parasitics, you will find that there significant parasitics associated with the PCB. When you solder a component in series with a lead, the placement of the component relative to the ground plane will affect the inductance and capacitance of the component. Likewise, when you solder a component to ground, the effect of the via path will affect the inductance. Traces between components can be modeled as *LC* circuits at low frequencies if the length of the trace

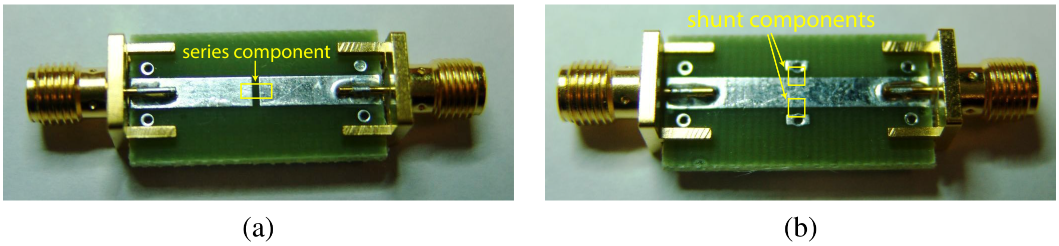
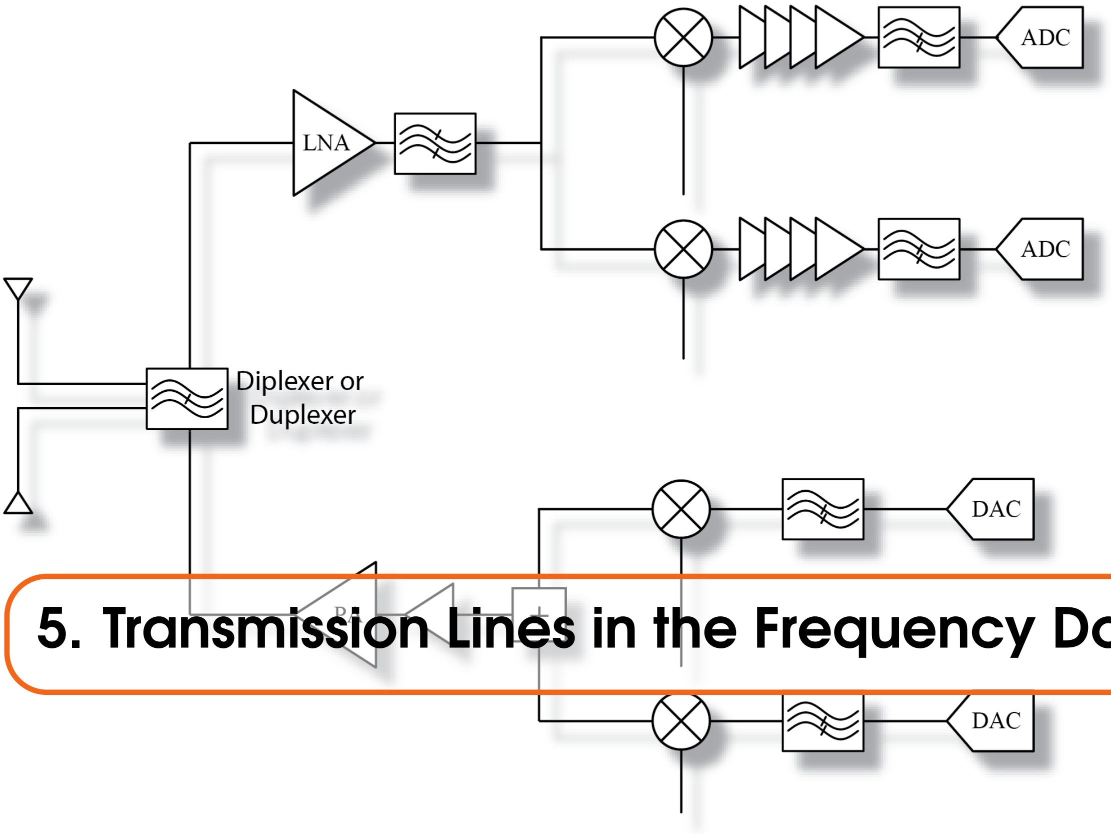


Figure 4.30: Structures used to estimate the board level parasitics. (a) An "open" structure allows a component to be soldered in series and characterized. (b) Shunt components can be soldered to ground to characterize components soldered in this configuration. In each case a zero ohm resistor can be used to estimate the series/shunt inductance.

is much shorter than the wavelength $\ell \ll \lambda$. For example, ideally a short circuit should have zero impedance but as the measurements will show, there is a finite amount of inductance and resistance below the self-resonant frequency. When a component is soldered to ground, there is additional inductance and resistance associated with the via to the ground plane. It is important to realize that the ground plane itself contributes resistance, especially at higher frequencies when the current flow is non-uniform and flows in the proximity of the transmission lines. Furthermore, the "inductance" of the components is strongly related to the "return current", or the path of the current flow under the component. If the ground path beneath the component is interrupted, forcing the current to flow away from the component, the parasitic inductance increases considerably.

You can estimate the parasitics of a 0603 component due to the physical size of the structure by using the "port extension" feature of the network analyzer. The reference plane for the measurement can be moved from the SMA connectors to the point where the device is connected. To do this you will need to measure an "open" and "short" as shown in Fig. 4.29 (include the loss) and the network analyzer will automatically calculate the phase delay and loss associated with the input and output transmission lines. Using this configuration, you can measure the parasitics of a component in series or in shunt using the test boards shown in Fig. 4.30. For instance, you can measure the two-port parameters of a zero-ohm resistor when in series or shunt with the signal path. The shunt components should be measured separately to include the effect of the via ground. Using the parasitics of a zero-ohm resistor allows you to modify your schematic to estimate the effect of the component parasitics. Alternatively, you can also measure the actual performance of the lumped components (inductors, capacitors) and model the component loss and resonance with an appropriate equivalent circuit.



5. Transmission Lines in the Frequency Domain

5.0.1 Telegrapher's Time Harmonic Equations

Now we are in position to derive the famous Telegrapher's Equations. To do this, consider the infinitesimal section of length δz of the line at an arbitrary distance z from the input. As shown in Fig. 5.1, the voltage at the output of the section is related to the input by

$$v(z+\delta z) = v(z) - i(z)Z_1'\delta z \tag{5.1}$$

Similarly the currents are related by

$$i(z) = i(z+\delta z) + \delta z Y_2' v(z+\delta z) \tag{5.2}$$

These equations can be converted into differential form if we take the limit of $\delta z \to 0$

$$\frac{dv}{dz} = -Z_1' i(z) \tag{5.3}$$

$$\frac{di}{dz} = -Y_2' v(z) \tag{5.4}$$

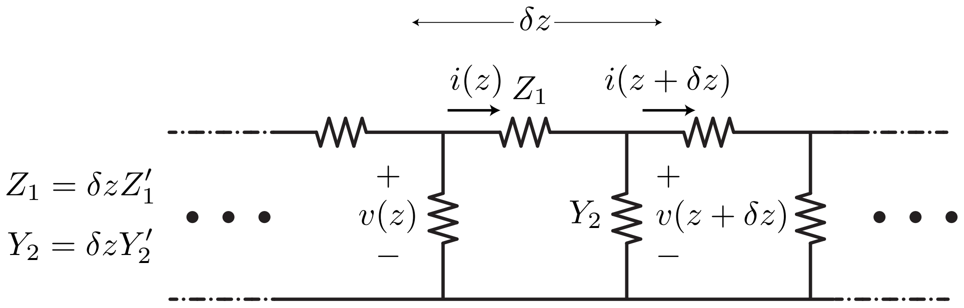


Figure 5.1: An infinitesimal section of the distributed transmission line.

These equations can be combined by taking the derivation of Eq. 5.3 and substituting from Eq. 5.4

$$\frac{d^2v}{dz^2} = -Z_1' \frac{di}{dz} = Z_1' Y_2' v(z) \tag{5.5}$$

The same equation applies to the current. Let the constant $Z_1' Y_2'$ be denoted as γ^2, resulting in the following set of boundary value differential equations

$$\frac{d^2v}{dz^2} = \gamma^2 v(z) \tag{5.6}$$

$$\frac{d^2i}{dz^2} = \gamma^2 i(z) \tag{5.7}$$

The general solution to the above equation is a complex exponential function

$$g(x) = G^+ e^{-\gamma z} + G^- e^{\gamma z} \tag{5.8}$$

the G^+ term is known as the "forward" traveling wave and the G^- term is known as the "reverse" or "backward" traveling wave. This is because when the time harmonic solution is written explicitly as a function of time and position

$$g(x,t) \propto e^{j(\omega t \pm \gamma z)} \tag{5.9}$$

which represents a propagating sinusoidal wave function in the forward and reverse z-direction. The propagation constant γ is in general complex

$$\gamma = \alpha + j\beta \tag{5.10}$$

We can thus factor the wave propagation function into a constant amplitude wave propagation multiplied by an envelope decay or growth

$$g(x,t) \propto e^{\pm \alpha z} e^{j(\omega t \pm \beta z)} \tag{5.11}$$

The phase of the complex exponential $\theta = \omega t \pm \beta z$ changes as a function of time

$$\frac{d\theta}{dt} = \omega \pm \beta \frac{dz}{dt} \tag{5.12}$$

To find the speed of the wave, we can follow a point on the wavefront. Such a point has constant phase θ, we set the above derivative equal to zero

$$\frac{d\theta}{dt} = 0 \tag{5.13}$$

and calculate the phase velocity v_p of the wave by noting how the position z must change in order to satisfy this condition. We find that

$$v_p = \frac{dz}{dt} = \pm \frac{\omega}{\beta} \tag{5.14}$$

Returning to the Telegrapher's Equation, we may write the general solution in the following form

$$v(z) = V^+ e^{-\gamma z} + V^- e^{\gamma z} \tag{5.15}$$

and

$$i(z) = I^+ e^{-\gamma z} + I^- e^{\gamma z} \tag{5.16}$$

An arbitrary transmission line will be denoted as shown in Fig. 5.2, an electrical representation for an arbitrary two conductor structure. At a given frequency, the line can be characterized by two complex numbers, the characteristic impedance Z_0 and the propagation constant γ. For the lossless line, there are only two relevant numbers, Z_0 and β, for all frequencies of propagation.

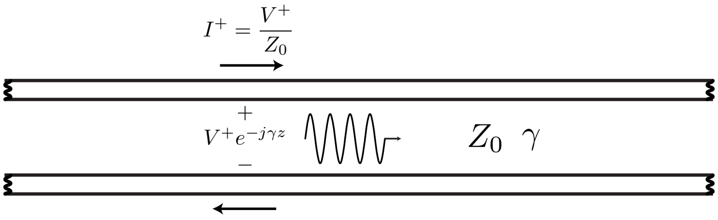


Figure 5.2: An electrical representation for a transmission line of characteristic impedance Z_0 and propagation constant γ.

5.0.2 Transmission Line Properties

The coefficients V^+, V^-, I^+, and I^- are related in a simple manner. By taking the derivative of the above equations and substituting the original relations from Eq. 5.3 and Eq. 5.4

$$-V^+\gamma e^{-\gamma z} + V^-\gamma e^{\gamma z} = -Z_1'\left(I^+e^{-\gamma z} + I^-e^{\gamma z}\right) \tag{5.17}$$

$$-I^+\gamma e^{-\gamma z} + I^-\gamma e^{\gamma z} = -Y_2'\left(V^+e^{-\gamma z} + V^-e^{\gamma z}\right) \tag{5.18}$$

These equations are satisfied for all values of z, and in particular for $z = 0$

$$-V^+\gamma + V^-\gamma = -Z_1'(I^+ + I^-) \tag{5.19}$$

$$-I^+\gamma + I^-\gamma = -Y_2'(V^+ + V^-) \tag{5.20}$$

which shows that only two of the four coefficients can possibly be independent. Let us define the impedance $Z_0 = \sqrt{Z_1'/Y_2'}$ and recall that $\gamma = \sqrt{Z_1'Y_2'}$. Then we can solve for I^+ and I^- in terms of V^+ and V^- to arrive at the following classic equations

$$v(z) = V^+e^{-\gamma z} + V^-e^{\gamma z} \tag{5.21}$$

$$i(z) = \frac{V^+}{Z_0}e^{-\gamma z} - \frac{V^-}{Z_0}e^{\gamma z} \tag{5.22}$$

In these equations the individual forward and backward waves begin to take on a life of their own. Notice that the current waves are related to the voltage waves by the characteristic impedance of the line Z_0. The backward wave, though, has a minus sign associated with it

$$Z_0^- = -\frac{V^-}{I^-} \tag{5.23}$$

A common case is the lossless line with $Z_1 = j\omega L$ and $Y_2 = j\omega C$. The characteristic impedance of the line is thus

$$Z_0 = \sqrt{\frac{j\omega L}{j\omega C}} = \sqrt{\frac{L}{C}} \tag{5.24}$$

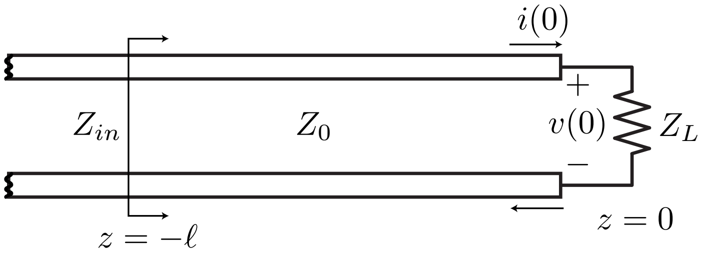


Figure 5.3: A transmission line terminated in a load impedance Z_L.

and the propagation constant is

$$\gamma = j\sqrt{LC}\omega \tag{5.25}$$

which is purely imaginary. The propagation velocity is given by

$$v = \pm\frac{\omega}{\beta} = \pm\sqrt{\frac{1}{LC}} \tag{5.26}$$

In Sec. 5.4, we shall show that this velocity is in fact the speed of light in the medium. For a two-wire transmission line suspended in air, this velocity is well known at approximately $3 \times 10^8\,\mathrm{m/s}$.

5.1 Transmission Line Termination

Up to now we have only considered an infinitely long uniform transmission in the z-direction. We found that such a structure supports voltages and currents which we have labeled as "forward" waves and "backward" waves (Eq. 5.21 and Eq. 5.22)

$$v(z) = V^+e^{-\gamma z} + V^-e^{\gamma z}$$

$$i(z) = \frac{V^+}{Z_0}e^{-\gamma z} - \frac{V^-}{Z_0}e^{\gamma z}$$

We would now like to focus on a terminated transmission line as shown in Fig. 5.3. In the figure we show a coaxial line terminated in a load impedance Z_L at $z=0$. Therefore, at the load the following relation must hold

$$Z_L = \frac{v(0)}{i(0)} \tag{5.27}$$

or substituting $z=0$ into the above equations

$$\frac{V^+ + V^-}{V^+ - V^-} = \frac{Z_L}{Z_0} \tag{5.28}$$

henceforth we shall denote normalized impedances such as $z_L = Z_L/Z_0$ with lowercase letters. Since there are two free variables V^+ and V^- and the constraint imposed by the load (Eq. 5.27), these parameters are now related. The variable $\rho_L = V^-/V^+$ parameterizes this relationship. For obvious reasons, we call ρ_L the load reflection coefficient since it represents the fraction of the wave "reflected" from the load relative to the forward wave. Thus rewriting the previous equation in our new notation, we have

$$z_L = \frac{1+\rho_L}{1-\rho_L} \tag{5.29}$$

The above equation is easily inverted

$$\rho_L = \frac{z_L - 1}{z_L + 1} \tag{5.30}$$

We see that when a transmission line is terminated, the voltage and current on the line take on more specific forms due to the constraint of Eq. 5.27 which results in a definite relationship between the "forward" and "backward" waves. Note in particular that $\rho_L = 0$ if $z_L = 1$, or $Z_L = Z_0$. This is the so called "matched line" load impedance that results in zero reflections. It's importance will be doubly highlighted when we consider transients on the transmission line. For a short circuit termination, $z_L = 0$ implies that $\rho_L = -1$, or in other words, the reflected backward wave has equal magnitude and opposite phase from the forward wave. This makes sense intuitively since the short circuit requires that $v(z = 0) = 0$, or the reflected wave must interfere destructively with the forward wave.

The current reflection coefficient is very simply related to the voltage reflected coefficient

$$\rho_{L,i} = \frac{I^-}{I^+} = \frac{-V^-/Z_0}{V^+/Z_0} = -\frac{V^-}{V^+} = -\rho_{L,v} \tag{5.31}$$

Thus for an an open termination $z_L = \infty$, zero current at the load implies that $\rho_{L,i} = -1$, or equivalently $\rho_{L,v} = +1$. Thus the reflected voltage signal has equal magnitude and phase from an open circuit.

5.2 Lossless Transmission Lines

The lossless transmission line is an important idealization for many everyday computations. Consider the arbitrary transmission line shown in Fig. 5.3 terminated in a load impedance Z_L, at some point $z = -\ell$ from the load. Since $\gamma = j\beta$ and $\alpha = 0$ for a lossless line, we may write Eq. 5.21 and 5.22 as

$$v(z) = V^+ e^{-j\beta z} + V^- e^{j\beta z} \tag{5.32}$$

$$i(z) = \frac{V^+}{Z_0} e^{-j\beta z} - \frac{V^-}{Z_0} e^{j\beta z} \tag{5.33}$$

Since the line is lossless, the propagation constant $\gamma = j\beta$ is imaginary and the characteristic impedance Z_0 is real. The constraint of the load impedance creates a reflected wave as before

$$\rho_L = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{5.34}$$

and in terms of ρ_L

$$v(z) = V^+ (e^{-j\beta z} + \rho_L e^{j\beta z}) \tag{5.35}$$

$$i(z) = \frac{V^+}{Z_0} \left(e^{-j\beta z} - \rho_L e^{j\beta z} \right) \tag{5.36}$$

The average power flow into the transmission line at an arbitrary point on the line is calculated by

$$P_{av} = \frac{1}{2} \Re \left[v(z) i(z)^* \right] \tag{5.37}$$

which can be written as

$$P_{av} = \frac{1}{2}\Re\left[\left(e^{-j\beta z} + \rho_L e^{j\beta z}\right)\left(e^{j\beta z} - \rho_L^* e^{-j\beta z}\right)\right] \times \frac{|V^+|^2}{Z_0} \tag{5.38}$$

and expanded

$$P_{av} = \Re\left[1 + \rho_L e^{2j\beta z} - \rho_L^* e^{-2j\beta z} - |\rho_L|^2\right] \times \frac{|V^+|^2}{Z_0} \tag{5.39}$$

notice that the middle terms in the brackets sum to a purely imaginary number [1], simplifying the calculations to

$$P_{av} = \frac{|V^+|^2}{2Z_0}\left(1 - |\rho_L|^2\right) \tag{5.40}$$

which is a constant independent of position. In particular, it's equal to the power delivered to the load. This of course follows from the lossless property of the line. Even though the power flow is constant on the line, the amplitude of the voltage and current waveforms are *not* constant unless $\rho_L = 0$.

5.2.1 Voltage Standing Wave Ratio (VSWR)

When the termination is matched to the line impedance $Z_L = Z_0$, $\rho_L = 0$ and thus the voltage along the line $|v(z)| = |V^+|$ is constant. Otherwise

$$|v(z)| = |V^+||1 + \rho_L e^{2j\beta z}| = |V^+||1 + \rho_L e^{-2j\beta\ell}| \tag{5.41}$$

where as before, ℓ is the distance away from the load. The reflection coefficient is in general a complex number

$$\rho_L = |\rho_L|e^{j\theta} \tag{5.42}$$

So the magnitude of voltage along the line can be written as

$$|v(-\ell)| = |V^+||1 + |\rho_L|e^{j(\theta-2\beta\ell)}| \tag{5.43}$$

The voltage is maximum when the $2\beta\ell$ is equal to $\theta + 2k\pi$, for any integer k; in other words, the reflection coefficient phase modulo 2π

$$v_{max} = |V^+|(1 + |\rho_L|) \tag{5.44}$$

and similarly, minimum when $\theta + k\pi$, where k is an integer $k \neq 0$

$$v_{min} = |V^+|(1 - |\rho_L|) \tag{5.45}$$

The ratio of the maximum voltage to minimum voltage is an important metric and commonly known as the voltage standing wave ratio, VSWR[2]

$$VSWR = \frac{v_{max}}{v_{min}} = \frac{1 + |\rho_L|}{1 - |\rho_L|} \tag{5.46}$$

[1] Since for any complex number a, $a - a^* = 2j\Im(a)$

[2] Sometimes pronounced viswar.

It's easy to show that the current standing wave ratio is the same as the voltage case, so it's also common to call the ratio the standing wave ratio, or SWR. It follows that for a shorted or open transmission line the VSWR is infinite, since $|\rho_L| = 1$.

Physically the maxima occur when the reflected wave adds in phase with the incoming wave, and minima occur when destructive interference takes place. The distance between maxima and minima is 180○ in phase, or $2\beta\Delta z = \pi$, or

$$\Delta z = \frac{\pi}{2\beta} = \frac{\lambda}{4} \tag{5.47}$$

VSWR is an important concept because it can be deduced with a relative measurement. Absolute measurements of impedance are difficult and impractical at microwave frequencies and a VSWR measurement allows one to deduce the impedance of an arbitrary load terminating a transmission line. By measuring VSWR, we can readily calculate $|\rho_L|$ by Eq. 5.1. By measuring the location of the voltage minima from an unknown load, we can solve for the load reflection coefficient phase θ

$$\psi_{min} = \theta - 2\beta\ell_{min} = \pi \tag{5.48}$$

Thus an unknown impedance can be characterized at microwave frequencies by measuring VSWR and ℓ_{min} and computing the load reflection coefficient. This important measurement technique that has been largely supplanted by a modern network analyzer with built-in digital calibration and correction.

Example 3:

Consider a transmission line terminated in a load impedance $Z_L = (1+2j)Z_0$. The reflection coefficient at the load is given by

$$\rho_L = \frac{z_L - 1}{z_L + 1} = \frac{1+2j-1}{1+2j+1} = \frac{j}{1+j} = \sqrt{\frac{1}{2}} e^{j\frac{\pi}{4}}$$

Since $1 + |\rho_L| = (\sqrt{2}+1)/\sqrt{2}$ and $1 - |\rho_L| = (\sqrt{2}-1)/\sqrt{2}$, the VSWR is equal to $\frac{\sqrt{2}+1}{\sqrt{2}-1}$. A plot of the voltage and current along a transmission line is shown in Fig. 5.4. Notice the location of the voltage maxima occurs when the reflection coefficient $|\rho_L| e^{j\frac{\pi}{4}} e^{-j2\beta\ell}$ is a real number, which occurs when $2\beta\ell = \frac{\pi}{4}$, or $\ell/\lambda = \frac{1}{16}$.

5.2.2 Transmission Line Input Impedance

We have seen that the voltage and current along a transmission line are altered by the presence of a load termination. At an arbitrary point z shown in Fig. 5.3, we wish to calculate the input impedance, or the ratio of the voltage to current as a function of Z_L

$$Z_{in}(-\ell) = \frac{v(-\ell)}{i(-\ell)} \tag{5.49}$$

It shall be convenient to define an analogous reflection coefficient at an arbitrary position along the line

$$\rho(-\ell) = \frac{V^- e^{-j\beta\ell}}{V^+ e^{j\beta\ell}} = \rho_L e^{-2j\beta\ell} \tag{5.50}$$

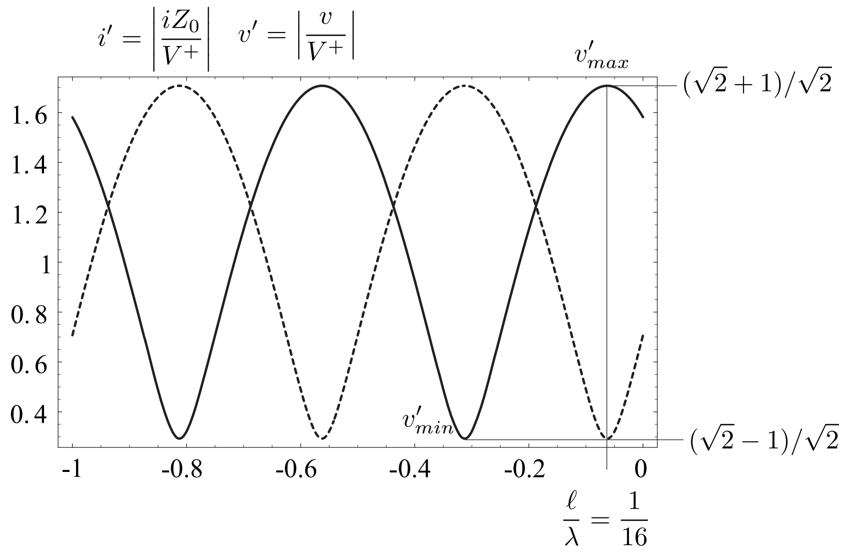


Figure 5.4: The magnitude of the steady-state time harmonic current (dotted) and voltage (solid) waveforms along a transmission line with $z_L = (1+2j)$.

which has a constant magnitude equal to the reflection coefficient at the load but a periodic phase. From this we may infer that the input impedance of a transmission line is also periodic since the relation between reflection coefficient and impedance is one-to-one

$$Z_{in}(-\ell) = \frac{v(-\ell)}{i(-\ell)} = \frac{V^+(1+\rho_L e^{-2j\beta\ell})}{V^+(1-\rho_L e^{-2j\beta\ell})} Z_0 \tag{5.51}$$

or eliminating the common voltage

$$Z_{in}(-\ell) = Z_0 \frac{1+\rho_L e^{-2j\beta\ell}}{1-\rho_L e^{-2j\beta\ell}} \tag{5.52}$$

The above equation is of paramount importance as it expresses the input impedance of a transmission line as a function of position ℓ away from the termination. This equation can be transformed into another more useful form by substituting the value of ρ_L

$$\rho_L = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{5.53}$$

and collecting terms

$$Z_{in}(-\ell) = Z_0 \frac{Z_L(1+e^{-2j\beta\ell}) + Z_0(1-e^{-2j\beta\ell})}{Z_0(1+e^{-2j\beta\ell}) + Z_L(1-e^{-2j\beta\ell})} \tag{5.54}$$

Using the common complex expansions for sine and cosine

$$\tan(x) = \frac{\sin(x)}{\cos(x)} = \frac{(e^{jx} - e^{-jx})/2j}{(e^{jx} + e^{-jx})/2} \tag{5.55}$$

which allows Eq. 5.54 to be replaced by the most important equation of the chapter, known sometimes as the "transmission line equation"

$$Z_{in}(-\ell) = Z_0 \frac{Z_L + jZ_0\tan(\beta\ell)}{Z_0 + jZ_L\tan(\beta\ell)} \tag{5.56}$$

Several special cases warrant individual attention.

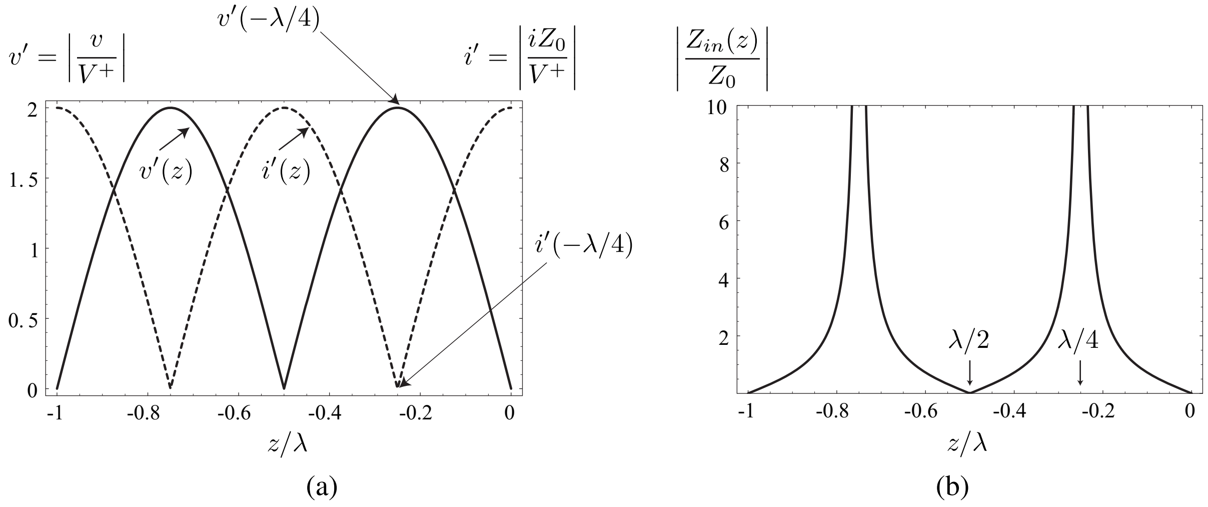


Figure 5.5: (a) The normalized magnitude of the current and voltage along a short-circuited transmission line as a function of position. (b) The magnitude of the impedance of a short circuited transmission line as a function of position.

Shorted Transmission Line

As already noted, the shorted transmission line has infinite VSWR and $\rho_L = -1$. Thus the minimum voltage $v_{min} = |V^+|(1 - |\rho_L|) = 0$, as expected. At any given point along the transmission line

$$v(z) = V^+(e^{-j\beta z} - e^{j\beta z}) = -2jV^+\sin(\beta z) \tag{5.57}$$

whereas the current is given by

$$i(z) = \frac{V^+}{Z_0}(e^{-j\beta z} + e^{j\beta z}) \tag{5.58}$$

or

$$i(z) = \frac{2V^+}{Z_0}\cos(\beta z) \tag{5.59}$$

and so the impedance at any point along the line takes on a simple form

$$Z_{in}(-\ell) = \frac{v(-\ell)}{i(-\ell)} = jZ_0\tan(\beta\ell) \tag{5.60}$$

which is just a special case of the more general transmission line equation with $Z_L = 0$. In particular note that the impedance is purely imaginary since a short and a lossless transmission line cannot dissipate any power. The line, though, stores reactive power in a distributed fashion. A plot of the normalized voltage and current, and input impedance as a function of z is shown in Fig. 5.5a,b.

Since the tangent function takes on infinite values when $\beta\ell$ approaches $\pi/2$ modulo 2π, the shorted transmission line can have infinite input reactance! This is particularly surprising since the load is in effect transformed from a short of $Z_L = 0$ to an infinite impedance.

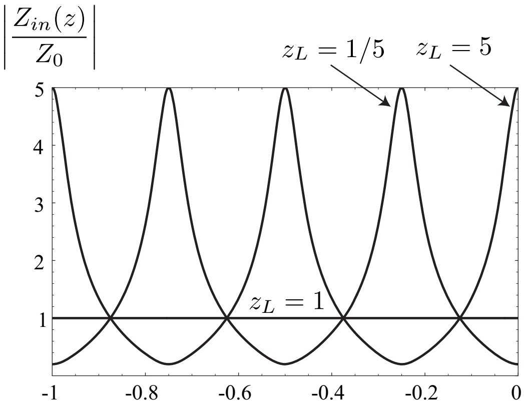


Figure 5.6: The magnitude of the impedance as a function of position on the transmission line for $z_L = 1$, $z_L = 5$ and $z_L = 1/5$.

Example 4: A transmission line with $Z_0 = 100\Omega$ is terminated in a load impedance of either $Z_L = 500\Omega$ or $Z_L = 20\Omega$. Can a VSWR measurement be used to determine the load?

Observe that the VSWR can be written as

$$VSWR = \frac{1+|\rho_L|}{1-|\rho_L|} = \frac{|1+z_L|+|1-z_L|}{|1+z_L|-|1-z_L|}$$

which can be simplified to

$$VSWR = \begin{cases} \frac{1}{z_L} & \text{if} \quad z_L < 1 \\ z_L & \text{if} \quad z_L > 1 \end{cases}$$

Thus the two cases above both result in a $VSWR = 5$. A plot of the normalized voltage is shown in Fig. 5.6. It is clear that the maxima occurs first on the transmission line with $z_L = 5$. This is easily remembered using a Smith Chart (see section 6.2).

5.3 Lossy Transmission Lines

We can account for loss in a transmission line by working directly with Eq. 5.21 and Eq. 5.22. In particular, all the calculation of the previous section can be redone to yield similar expressions.

5.3.1 Transmission Line Input Impedance

Taking the ratio of voltage to current on the transmission line, we have the general expression

$$Z_{in}(z) = \frac{v(z)}{i(z)} = Z_0 \frac{V^+e^{-j\gamma z}+V^-e^{j\gamma z}}{V^+e^{-j\gamma z}-V^-e^{j\gamma z}} \tag{5.61}$$

Using the definition of the reflection position, $\rho(z)$, we have

$$Z_{in}(z) = \frac{1+\rho(z)}{1-\rho(z)} \tag{5.62}$$

which can be written explicitly in terms of the load impedance

$$Z_{in}(-\ell) = Z_0 \frac{Z_L + Z_0 \tanh(\gamma\ell)}{Z_0 + Z_L \tanh(\gamma\ell)} \tag{5.63}$$

It's easy to show that the above equation degenerates to Eq. 5.2.2 under the lossless case.

Example 5:

Calculate the input impedance of a short-circuited transmission line of length ℓ at low frequency. Assume the line is very short, e.g. $\gamma\ell \approx 0$. From Eq. 5.3.1, substituting $Z_L = 0$, we have

$$Z_{in}(-\ell) = Z_0 \tanh(\gamma\ell) \approx Z_0\gamma\ell$$

Recalling the definition of γ and Z_0, we have

$$Z_{in}(-\ell) \approx \sqrt{\frac{Z_1'}{Y_2'}}\sqrt{Z_1'Y_2'}\ell = Z_1'\ell = R_T + j\omega L_T$$

where R_T is the resistance of the line and L_T is the inductance of the line. Thus a shorted transmission line behaves like a lumped inductor.

5.3.2 Dispersionless Line

In general, a lossy transmission line introduces distortion due to dispersion. Dispersion occurs when the propagation speed and attenuation is frequency dependent. If a group of frequencies are excited along the line, they travel along the line with different velocity and experience different attenuation. Thus, if we an arbitrary waveform (say a pulse) is excited on the line, after significant propagation, it will arrive with a completely distorted waveform since it's different frequency components will be time shifted and attenuated unevenly.

For a dispersionless line, the output should be a linearly scaled delayed version of the input $v_{out}(t) = Kv_{in}(t-\tau)$, or in the frequency domain

$$V_{out}(j\omega) = KV_{in}(j\omega)e^{-j\omega\tau} \tag{5.64}$$

The transfer function has constant magnitude $|H(j\omega)|$ and linear phase $\angle H(j\omega) = -\omega\tau$. The propagation constant $j\beta$ should therefore be a linear function of frequency and α should be a constant. We can find the conditions for the transmission line to be dispersionless in terms of the R, L, C, G, expand

$$\gamma = \sqrt{(j\omega L' + R')(j\omega C' + G')} \tag{5.65}$$

$$= \sqrt{(j\omega)^2 LC(1 + \frac{R}{j\omega L} + \frac{G}{j\omega C} + \frac{RG}{(j\omega)^2 LC})} \tag{5.66}$$

$$= \sqrt{(j\omega)^2 LC}\sqrt{\square} \tag{5.67}$$

Suppose that $R/L = G/C$ and simplify the $\square$ term

$$\square = 1 + \frac{2R}{j\omega L} + \frac{R^2}{(j\omega)^2 L^2} \tag{5.68}$$

For $R/L = G/C$ the propagation constant simplifies

$$\square = \left(1 + \frac{R}{j\omega L}\right)^2 \tag{5.69}$$

$$\gamma = -j\omega\sqrt{LC}\left(1 + \frac{R}{j\omega L}\right) \tag{5.70}$$

Breaking γ into real and imaginary components

$$\gamma = R\sqrt{\frac{C}{L}} - j\omega\sqrt{LC} = \alpha + j\beta \tag{5.71}$$

The attenuation constant α is independent of frequency. For low loss lines, $\alpha \approx -\frac{R}{Z_0}$. The propagation constant β is a linear function of frequency.

Therefore, to design a dispersionless lossy line, we must strive to equalize R/L and G/C. One way to achieve this is to periodically load the line with a lumped capacitor or inductor to force equality. For instance, if series lumped inductors of value L_x are used, this forms an *artificial* transmission line with inductance per unit length given by $L'' = L' + L_x/d$, if the distance d between elements is much smaller than the wavelength of propagation. Then the line behaves very much like an ideal transmission line for frequencies up to a cutoff frequency.

5.3.3 Power Flow on a Lossy Line

It's interesting to calculate the steady-state power delivered to a load termination of a transmission line

$$P_L = \frac{1}{2}\Re\left[V_L I_L^*\right] \tag{5.72}$$

Expanding the voltage and current at the load in terms of of the forward and backwards waves, we can rewrite the above as

$$P_L = \frac{1}{2}\Re\left[(V^+ + V^-)(V^{+*} - V^{-*})\frac{1}{Z_0^*}\right] \tag{5.73}$$

which can be rewritten in the following form

$$P_L = \frac{1}{2}|V^+|^2\Re(Y_0)(1 - |\rho_L|^2) \tag{5.74}$$

The first term can be interpreted to represent the incident power of the forward wave, which implies that the second term represents the power reflected from the load. Clearly, maximum power delivery to the load occurs when $\rho_L = 0$, or when the load is matched to the transmission line impedance Z_0.

For a lossy transmission line, the power delivered into the line at a point z is non-constant and decaying exponentially.

$$P_{av}(z) = \frac{1}{2}\Re\left(v(z)i(z)^*\right) = \frac{|v^+|^2}{2|Z_0|^2}e^{-2\alpha z}\Re\left(Z_0\right) \tag{5.75}$$

Compare the above equation to Eq. 5.2, where all the power injected into a lossless line flows to the load. In the lossy case, more power is required due to attenuation constant α.

For instance, if $\alpha = .01\text{m}^{-1}$, then a transmission line of length $\ell = 10\text{m}$ will attenuate the signal by $10\log(e^{2\alpha\ell})$ or 2 dB. At $\ell = 100\text{m}$, the line attenuates the signal by $10\log(e^{2\alpha\ell})$ or 20 dB. The attenuation constant α plays a very important role since it essentially determines the maximum length of a transmission line before requiring signal amplification. If the signal is attenuated too much, it will be buried in the natural noise of the system.

If the load is mismatched, then we must consider the reflected waves

$$P_{in} = \frac{1}{2}\Re\left(v(-\ell)i(-\ell)^*\right) \tag{5.76}$$

$$= \frac{1}{2}\frac{|V^+|^2}{Z_0}\Re\left[\left(e^{\gamma\ell} + \rho e^{-\gamma\ell}\right)\left(e^{\gamma^*\ell} - \rho^* e^{-\gamma^*\ell}\right)\right] \tag{5.77}$$

where we assumed a low loss line so that Z_0 is essentially a real quantity. Expanding the expression we have

$$= \frac{1}{2}\frac{|V^+|^2}{Z_0}\Re\left[e^{2\alpha\ell} + \underbrace{\rho e^{-2j\beta\ell} - \rho^* e^{2j\beta\ell}}_{\text{imaginary}} - |\rho|^2 e^{-2\alpha\ell}\right] \tag{5.78}$$

which simplifies to

$$P_{in} = \frac{|V^+|^2}{2Z_0}\Re\left[e^{2\alpha\ell} - |\rho|^2 e^{-2\alpha\ell}\right] \tag{5.79}$$

At the load we have

$$P_L = P_{in}(0) = \frac{|V^+|^2}{2Z_0}\Re\left(1 - |\rho|^2\right) \tag{5.80}$$

and so subtracting the above from Eq. 5.79 gives the power dissipated by the transmission line

$$P_{diss} = P_{in} - P_L = \frac{|V^+|^2}{2Z_0}\left[\left(e^{2\alpha\ell} - 1\right) + |\rho|^2\left(1 - e^{-2\alpha\ell}\right)\right] \tag{5.81}$$

We can interpret the first term as the power dissipated by the incoming wave and the second term as the power dissipated by the reflected wave.

5.4 Field Theory of Transmission Lines

In more general terms, a transmission line is a structure that supports transverse electromagnetic (TEM) wave propagation. By "transverse" we mean that the fields are always perpendicular to the direction of propagation. For a transmission line aligned with the z-axis, the fields cannot have a z-component. There are several important and simplifying consequences to this assumption. First, since currents flow into and out of the conductors of a transmission line, $J_z \neq 0$ but $E_z \equiv 0$. This implies that the conductors must be infinitely conductive, or lossless. Furthermore, since the axial magnetic field is zero, it follows that the E-field behaves statically in the transverse plane

$$\int_{\text{t-plane}} \mathbf{E}\cdot d\ell = -\frac{d}{dt}\int_{\text{t-plane}} \mathbf{B}\cdot d\mathbf{S} \equiv 0 \tag{5.82}$$

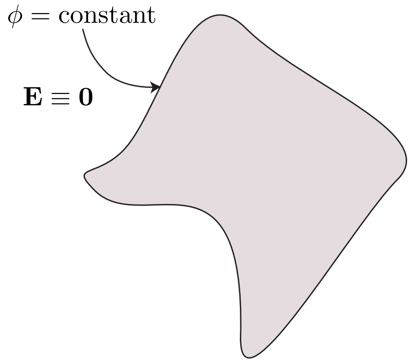


Figure 5.7: A single conductor cannot support TEM waves since the electrostatic solution has only a constant potential solution.

and we can thus define a unique voltage in the transverse plane

$$v(z,t) = -\int_1^2 \mathbf{E} \cdot d\ell \tag{5.83}$$

Similarly, since the axial electric field is zero, the magnetic field is solely dependent on the current

$$\int_{\text{t-plane}} \mathbf{H} \cdot d\ell = -\frac{d}{dt} \int_{\text{t-plane}} \mathbf{E} \cdot d\mathbf{S} + \mu I \tag{5.84}$$

Since the displacement current is identically zero, we define a unique current

$$\int_{\text{t-plane}} \mathbf{H} \cdot d\ell = \mu I \tag{5.85}$$

and the magnetic field also behaves statically.

In summary, for TEM propagation, the fields have zero axial component, $E_z \equiv H_z \equiv 0$, and the transverse components of the fields behave like static fields. Thus, for a uniform structure the laws of electromagnetics reduce to statics regardless of the frequency of excitation! While only lossless conductors can truly support TEM waves, in practice, low-loss conductors are often employed and behave essentially like TEM guides. Finally, it can be proved that TEM waves can only be supported by *two* or more conductors. Any single conductor cannot support TEM waves because a single ideal conductor cannot support a static field solution. For instance, consider the structures shown in Fig. 5.7. The solution to Poisson's equation, for example, is found by noting that the constant potential is a permissible solution, e.g. $\mathbf{E} \equiv \mathbf{0}$ everywhere including on the surface of the conductor. But since the solution to Poisson's equation is unique, this is the *only* solution. Thus more than one conductor is required for TEM modes. While such a conductor cannot support TEM waves, it can support transverse electric (TE) or transverse magnetic (TM) waves, or any combination thereof.

It's important to observe that an ideal transmission line has a constant characteristic impedance Z_0 versus frequency. Since the fields are the solution of the static fields in the transverse plane, there is no frequency dependence in the fields and thus no variation in the inductance and capacitance. This can be seen in another way. Since all conductors are perfectly conducting, no fields can penetrate the conductors and therefore only external inductance contributes to magnetic energy storage in the line. Thus the inductance is constant as a function of frequency. In practice, of course, all transmission lines have loss. We have already modeled the loss in our distributed circuits with an equivalent series and shunt resistor in the ladder network. So in practice, the transmission line

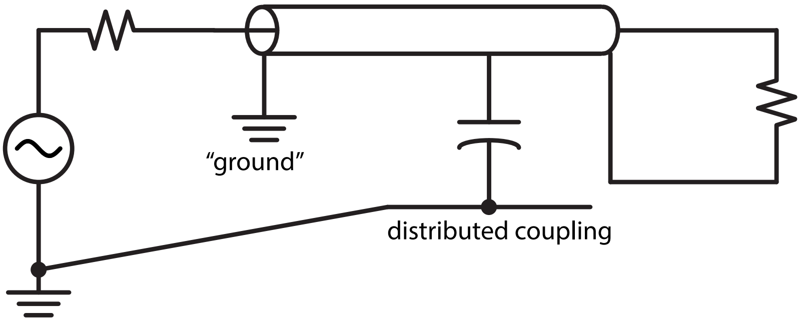


Figure 5.8: Consider a coaxial cable connected to the “ground", which may be the ground plane of the PCB or a metal frame or chassis. Even though the currents are routed to the shield at the load, and we would expect all the currents to flow inside the coax, due to distributed coupling to the ground, some current may return via this undesired path, introducing a new propagation mode.

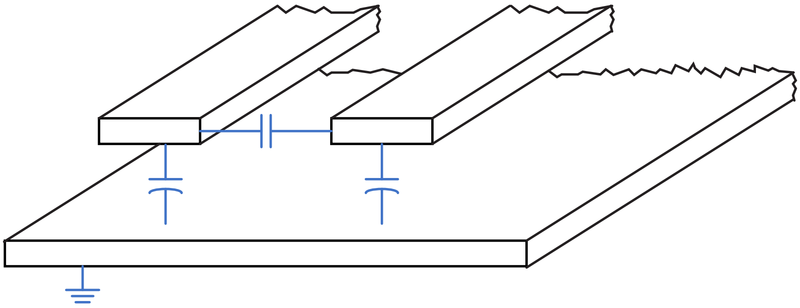


Figure 5.9: A stripline is used to route a differential signal, but due to the proximity of the ground plane, currents can also easily return through the ground plane, or stripline modes for each conductor.

properties do depend on frequency. While it is possible to account for the loss in our field equations, it’s much better to treat the loss as a perturbation to an ideal line rather than manipulate the full blown Maxwell’s equations from the outset.

5.5 Multimode Propagation

Up to now, we have assumed that the signal travels along two conductors, one “signal" conductor and one “ground" conductor. As shown in Fig. 5.8, even if we “ground" the second conductor, there’s no reason that it has to stay “grounded". In fact, like it or not, there’s always a third conductor at play. For example, when two wires are routed on a PCB, the ground plane can act as a potential signal return path (so can earth ground). So we really have to consider the possibility of exciting this “common mode" transmission line.

Stripline Example

In Fig. 5.9 we have two conductors that form a stripline (differential line) but they reside above a ground plane. So each one individually couples to the ground plane and forms another transmission line. In general, we must model this secondary transmission line to account properly for the actual signal propagation.

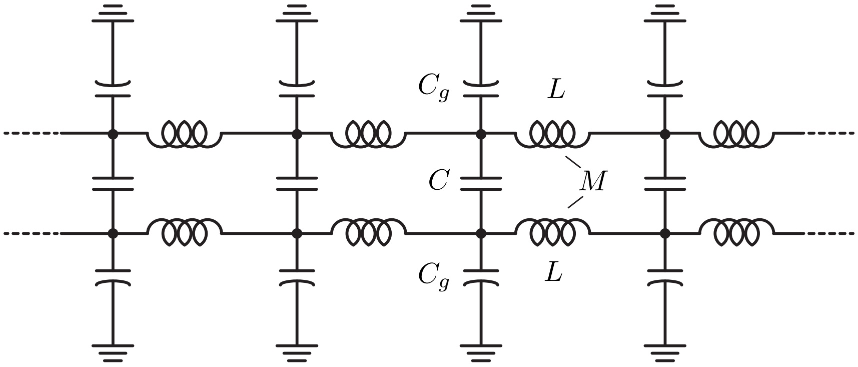


Figure 5.10: A sectional distributed model of two coupled lines. Coupling occurs through mutual capacitance C and mutual inductance M.

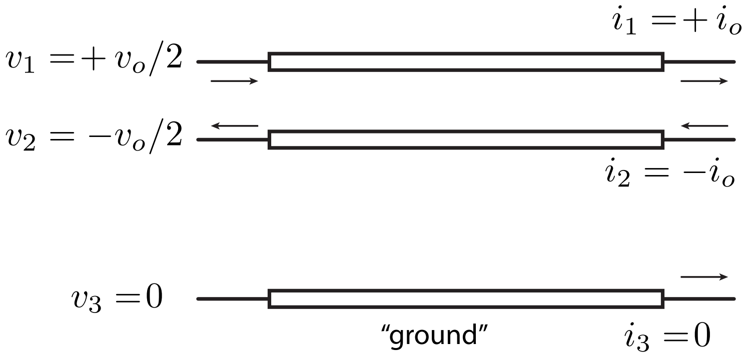


Figure 5.11: We can model our stripline using a multi-line model shown. Here we track voltages/currents on each line. Odd mode corresponds to $v_1 = -v_2$.

5.5.1 Distributed Multi-Line Model

In the sectional model shown in Fig. **??**, we show a distributed inductance and capacitance to the ground as before, but also mutual inductance and capacitance. Note that the return current flowing on the ground plane also contributes to the inductance (mutually coupled to all other conductors) but we can effectively lump the entire loop of signal plus ground into one inductance as shown. This is the same thing we did for the two conductor model.

Odd Mode Excitement

If we assume that $v_1 = -v_2$, we say we are exciting the odd mode. In this mode, no signal current flows into the ground plane as $i_1 = -i_2$, see Fig. 5.11. This mode can be excited in a symmetric structure with a differential circuit. Note that if we ground v_2 at the source, we are not exciting *only* the odd mode. This is an important and subtle point you should think about!

The propagation constant is given by

$$Z_o = \sqrt{\frac{L-M}{C+\frac{C_g}{2}}} \tag{5.86}$$

The minus sign for M is due to the fact that the currents in line 1 and 2 are in opposite phase. The lines couple through mutual capacitance and through the ground plane C_g.

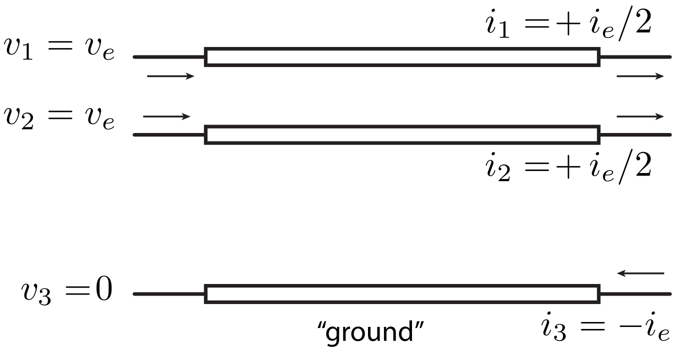


Figure 5.12: We model our stripline using a multi-line model shown. Even mode corresponds to $v_1 = v_2$.

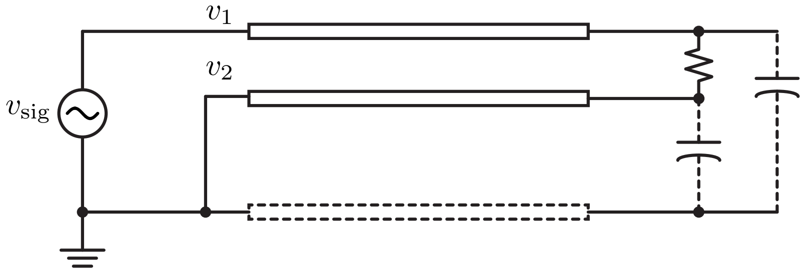


Figure 5.13: A grounded voltage source drives a line that supports both even and odd modes.

Even Mode Excitement

If we assume that $v_1 = v_2$, we say we are exciting the even mode, shown in Fig. 5.12. In this mode, there is necessarily a signal current flowing into the ground plane. This mode can be excited by shorting the two signal conductors together and observing the current flow.

The propagation constant is given by

$$Z_e = \sqrt{\frac{L+M}{2C_g}} \tag{5.87}$$

The plus sign with M corresponds with the currents in the two conductors flowing in the same direction, whereas the mutual capacitance C is essentially open-circuited since $v_1 = v_2$ and no current flows through it.

5.5.2 General Modal Excitement

In general, we can excite both the even and odd modes. In the example shown in Fig. **??**, if we ground one side of the T-line and drive it as shown in a single-ended fashion, we are exciting both modes

$$v_1 = v_{sig} = v_e + v_o \tag{5.88}$$

$$v_2 = 0 = v_e - v_o \tag{5.89}$$

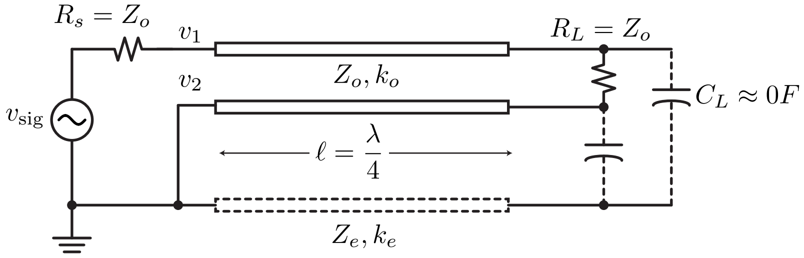


Figure 5.14: Over a narrow range in frequencies, a $\lambda/4$ line, or a quarter wave line, acts like a balun, converting a single-ended voltage drive into a balanced drive.

Note that both even and odd mode is excited in this case. The difference in voltage between the lines is the odd mode

$$2v_o = v_1 - v_2 \tag{5.90}$$

$$v_o = (v_1 - v_2)/2 = v_{sig}/2 \tag{5.91}$$

While the average voltage on the lines is the even mode

$$v_1 + v_2 = 2v_e \tag{5.92}$$

$$v_e = (v_1 + v_2)/2 = v_{sig}/2 \tag{5.93}$$

Since the above excites both the even and odd modes, we have to take them into account

$$v_1(z) = v_o(e^{jk_oz} + \rho_{L,o}e^{-jk_oz}) + v_e(e^{jk_ez} + \rho_{L,e}e^{-jk_ez}) \tag{5.94}$$

$$v_2(z) = -v_o(e^{jk_oz} + \rho_{L,o}e^{-jk_oz}) + v_e(e^{jk_ez} + \rho_{L,e}e^{-jk_ez}) \tag{5.95}$$

Where the current boundary conditions are captured by the load reflection coefficients $\rho_{e,o}$, which are not equal.

5.5.3 Example: Quarter Wave Balun

A transmission line that is $\lambda/4$ can be used as a balun, as shown in Fig. 5.14. To see this, let's solve both the even and odd modes for the case that the line is terminated. For the odd mode, the boundary condition is clearly determined by the load

$$\rho_o = \rho_L = \frac{Z_L - Z_o}{Z_L + Z_o} \approx 0 \tag{5.96}$$

Whereas for the even mode, let's say it's approximately an open circuit

$$\rho_e = \frac{Z_{open} - Z_e}{Z_{open} + Z_e} \approx 1 \tag{5.97}$$

Substituting in the above equations

$$v_1(z) = v_o e^{jk_o z} + v_e(e^{jk_e z} + e^{-jk_e z}) \tag{5.98}$$

$$v_2(z) = -v_o e^{jk_o z} + v_e(e^{jk_e z} + e^{-jk_e z}) \tag{5.99}$$

Notice that the second terms precisely cancel out at the load since

$$e^{j\pi/2} = +j \tag{5.100}$$

and

$$e^{-j\pi/2} = -j \tag{5.101}$$

This implies that the signal at the load side is a pure odd mode, or balanced signal. To achieve this, we desire a high common mode impedance to satisfy $\rho_e \approx 1$.

5.6 T-Line Structures

5.6.1 The Coaxial Line

The coaxial line shown in Fig. 2.7a is perhaps the most commonly encountered transmission line. In many residential areas, cable TV and Internet data services are delivered to homes via 75Ω transmission lines. Due to circular symmetry, the inductance and capacitance per unit length are readily calculated. An important observation comes from Eq. 5.26, which shows that the inductance and capacitance are in fact related. Thus, only one needs to be calculated and the other is easily inferred.

In Fig. 2.7a, the coaxial transmission line is shown with an inner conductor of radius a and outer conductor of radius b. One can easily calculate the inductance and capacitance per unit length for such a line

$$C' = \frac{2\pi\varepsilon}{\ln(b/a)} \tag{5.102}$$

since the propagation velocity of a TEM line is constant $v_p = 1/\sqrt{\varepsilon\mu} = 1/\sqrt{L'C'}$

$$L' = \frac{\varepsilon\mu}{C'} = \frac{\mu}{2\pi}\ln\left(\frac{b}{a}\right) \tag{5.103}$$

Keep in mind that this is the external inductance per unit length. Only a perfect conductor will carry all of its current on the outer skin. In the limit of very high frequency, though, or when the radius of the conductor is much larger than the skin depth, $a \gg \delta$, the above formula applies. The conductance per unit length due to a lossy dielectric is easily derived since $C'/G' = \varepsilon/\sigma$. The series resistance per unit length, though, is more difficult to calculate.

To minimize the loss of a coaxial transmission line, we should minimize the conductive and resistive losses. The conductive losses can be minimized by using a material with the lowest possible dielectric loss. Air is a pretty good insulator with virtually zero loss, vacuum is even better. But in practice a solid dielectric is preferred since it provides mechanical support.

Coaxial lines have several advantages. Since the outer conductor completely surrounds the inner conductor, no fields are to be found outside the structure that originate from charge and currents inside. This follows by application of Gauss' theorem and Ampere's law. From Gauss'

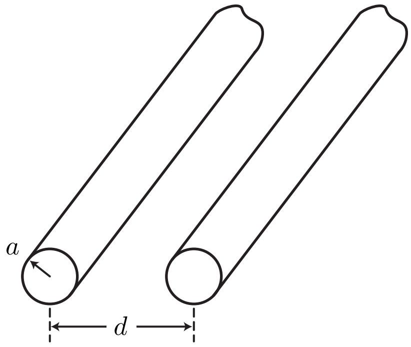


Figure 5.15: A balanced two-wire transmission line.

theorem, consider the volume of a cylinder completely surrounding the coaxial line. The surface integral of the electric field is proportional to the charge inside the volume

$$\int_{cylinder} \mathbf{E} \cdot d\mathbf{S} = \varepsilon Q_{\text{inside}} \tag{5.104}$$

The outer conductor is grounded and thus charges flow onto the inner surface of the outer conductor and effectively shield the conductor. For a properly grounded coaxial transmission line, the ground charge on the outer conductor is equal and opposite to the charge on the inner conductor. Therefore the surface integral is identically zero. But by symmetry the field is everywhere radial and of equal magnitude on the surface of the cylinder

$$\int_{cylinder} \mathbf{E} \cdot d\mathbf{S} = 2\pi R E_r \ell \tag{5.105}$$

where ℓ is the length of the cylinder. Thus the field is identically zero, $E_r \equiv 0$, everywhere outside of the coaxial transmission line. A similar argument can be made for the magnetic field implying that $H_t \equiv 0$ everywhere outside of the conductors.

5.6.2 Balanced Two-Wire Line

Consider the two wire transmission line shown in Fig. 5.15. The wires are separated by a distance d and each conductor has a radius of a. Each conductor carries an equal but opposite current. To find the magnetic field at a given point in space, notice that Ampère's Law does not help much since the second conductor ruins the symmetry of the problem. But an approximate formula for the inductance per unit length is easy to derive if we assume that $d \gg a$. In that case the currents in the wires do not interact and the total field is simply the sum of the magnetic field of each wire

$$B(r) = \frac{\mu I}{2\pi r_1} + \frac{\mu I}{2\pi r_1} \tag{5.106}$$

where r_1 and r_2 is the distance to the center of each conductor. Let's integrate the magnetic field along the plane containing the conductors to obtain the magnetic flux per unit length

$$\psi' = \int_a^{a+d} \frac{\mu I}{2\pi} \left(\frac{1}{x} + \frac{1}{d+2a-x} \right) dx \tag{5.107}$$

$$= \frac{\mu I}{\pi} \ln \frac{d+a}{a} \tag{5.108}$$

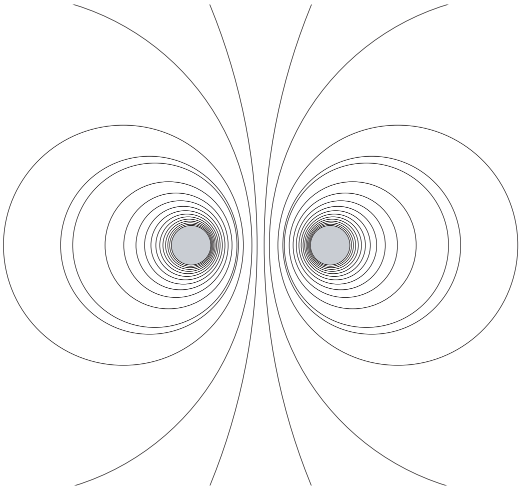


Figure 5.16: The contours of constant magnetic field for two filaments carrying opposite currents.

So that the inductance per unit length is simply

$$L' \approx \frac{\mu}{\pi} \ln \frac{d+a}{a} \tag{5.109}$$

An exact derivation will take some more work but the solution is considerably simplified if we simply consider the contours of constant magnetic field for two filamental currents with opposite polarity, as shown in Fig. 5.16. Notice that the contours are circles, allowing us to conveniently slip in two conductors with finite radius without changing the field distribution! Note that the filaments are *not* at the centers of the conductors. If you like to do algebra, then you can show that this is exactly true by plotting the contours by noting that on any given contour, the ratio of the distances to one filaments must be fixed, or $r_2/r_1 = k$

In the exact derivation, we find that

$$L' = \frac{\mu}{\pi} \cosh^{-1} \frac{d}{2a} \tag{5.110}$$

which agrees with our previous derivation for $d \gg a$.

The two wire transmission line is used as a low cost feedline for differential antennas. For instance, a UHF loop antenna of 300Ω is common, and a two-wire transmission line can be used to feed the structure. Note that most television sets use 75Ω transmission lines, so an impedance matching circuit is needed (otherwise you'd see a ghost on the TV screen). But since the matching network must work for all UHF channels, a narrowband matching network will not work. A broadband matching network using transmission line transformers will be discussed later.

5.6.3 Planar Differential Transmission Line

Since differential circuits are commonly used on-chip, a two-wire rectangular transmission line is very commonly used. This structure is also known as a coplanar stripline (CPS). As shown in Fig. 5.17, a differential pair can excite equal and opposite currents on a differential two-wire transmission line. In reality a pseudo-TEM wave propagates since the dielectric constant of the Si substrate differs from the oxide SiO_2, and thus any fields leaking into the substrate will travel at a slower velocity than the oxide. Such a structure cannot support a TEM wave, but the actual waves are a good approximation to TEM waves if the spacing between the conductors is not too large.

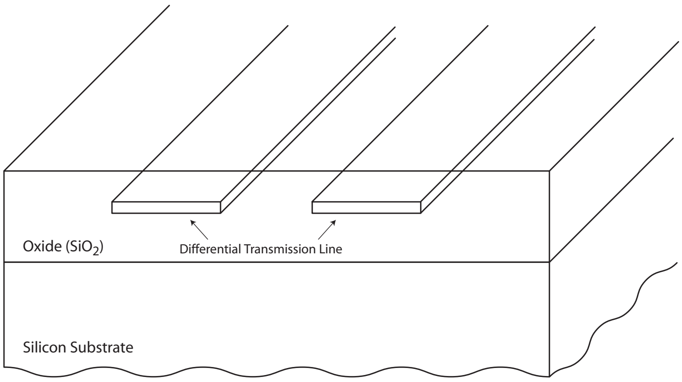


Figure 5.17: A balanced two-conductor transmission line implemented in an IC process.

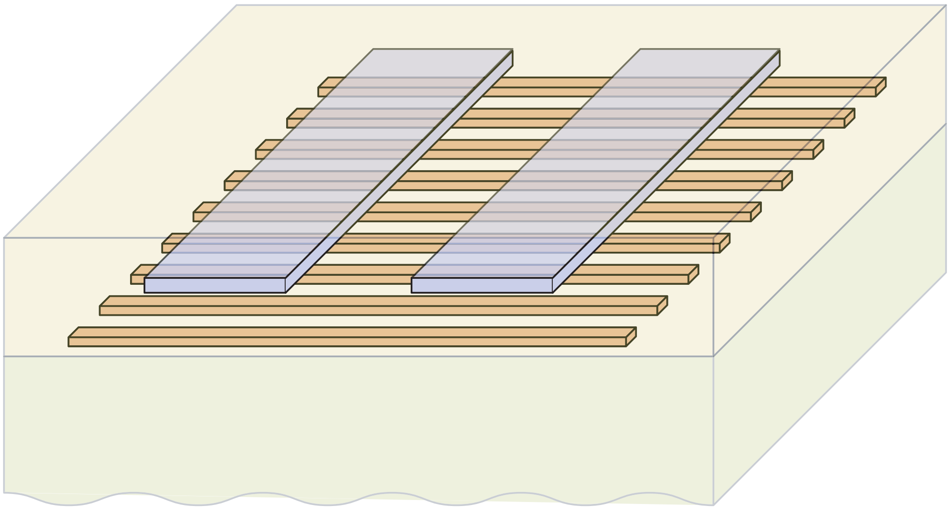


Figure 5.18: A balanced slow-wave two-conductor transmission line.

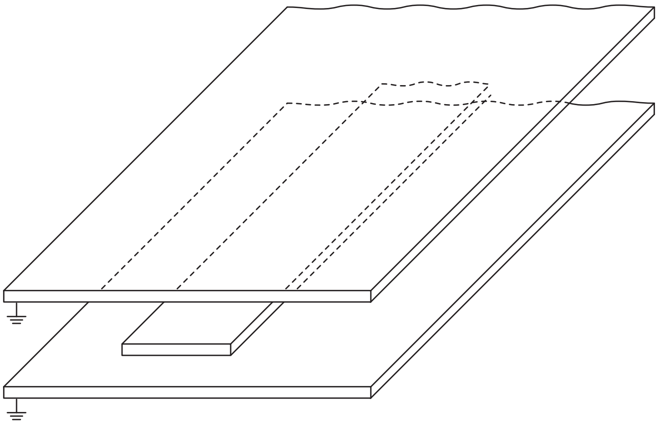


Figure 5.19: A three conductor stripline structure. The top and bottom conductors are both grounded

One way to avoid leakage into the substrate is to place a solid shield under the conductors. This lowers the characteristic impedance of the line since the inductance per unit length decreases while the capacitance per unit length increases. The wave velocity is practically unchanged, though, since the product LC is constant. We can lower the wave velocity with the slow wave structure shown in Fig. 5.18. Here, conductive strips are placed beneath the two wires to increase the capacitance per unit length without altering the inductance per unit length. The inductance is not lowered since current cannot flow perpendicular to the strips, and thus the transmission line current flow is confined to the differential conductors. Since L is unchanged while C increases, the product LC is increased corresponding to a lower phase velocity. Does this contradict our earlier claim that LC is a constant? No, because this structure is not a uniform transmission line. In fact, you can view this structure as an artificial dielectric with $\varepsilon' > \varepsilon$ due to the metal strips.

Even and Odd Modes

One important issue to be aware of is that the CPS structure can propagate waves in two modes, a “odd” mode, which is the desired differential mode, whereby the conductors are excited in a balanced out of phase fashion, and an “even" mode, also known as “common-mode” in circuits parlance (see Fig. **??**). In even mode both conductors are excited in phase and the return current flows from the ground (either the PCB ground or silicon substrate).

Since the odd mode is the desired mode, we should exercise caution to minimize exciting the even mode. Any imbalance in the driver or load can excite a common mode signal, which then launches an even mode wave. The propagation velocity and attenuation of the even mode is different, though, since the fields couple to the ground plane, which is partially conductive for silicon. By using two closely space lines, with a spacing smaller than the distance to the substrate, we can raise the even mode impedance. By carefully controlling the grounding of signals to the substrate, we also create a high impedance path so that most of the current favorably flows into the odd mode.

5.6.4 Planar Stripline and Microstrip Line

The stripline, shown in Fig. 5.19, is a planar version of the coaxial line, with a ground plane above and below the rectangular conductor. Such a structure can be constructed on a PCB or even on-chip. In practice the grounds are not infinite in extent, but wide enough to absorb all the fringing fields. The microstrip is a simpler structure, shown in Fig. 5.20, as one ground plane is eliminated. Since the fields fringe into the air on the top side, though, pure TEM waves cannot be supported. In practice most of the fields remain in the dielectric since $\varepsilon > \varepsilon_0$ and a TEM approximation is a good

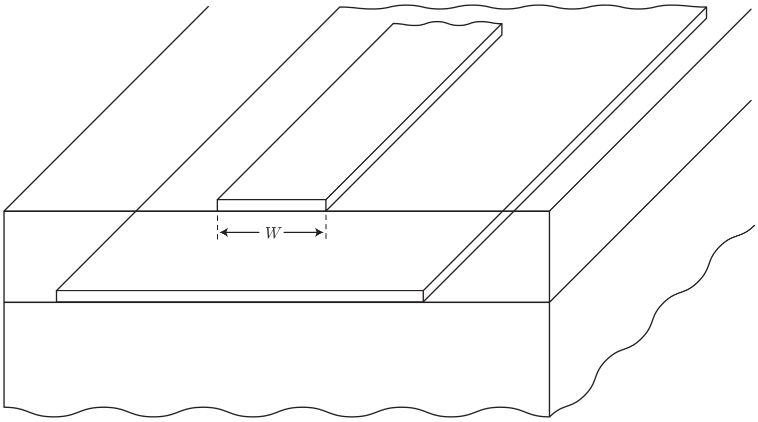


Figure 5.20: A microstrip transmission line.

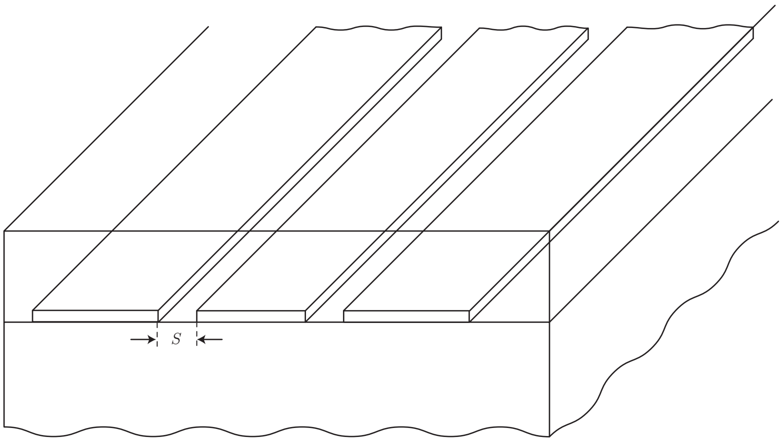


Figure 5.21: A coplanar transmission line.

one.

One advantage of a stripline over a microstrip is the loss per unit length. In a microstrip the ground current flows over two conductors, thus halving the return current losses. More importantly, the conductor current flows on the top and bottom faces, thus reducing the resistance per unit length. For the microstrip, the current flows primarily on the bottom face of the conductor.

5.6.5 Coplanar Lines

Microstrip lines have some disadvantages when implemented on-chip. Since the return current flows on the bottom metal, the spacing between the conductor and ground is fixed. Thus, the only way to control the line impedance is to change the conductor width W. For a high impedance line, we require small W to minimize C and maximize L. But this introduces more loss into the line. A coplanar line, shown in Fig. 5.21, uses ground return currents on the same metal layer. The line impedance is now a strong function of the spacing S. A high impedance line can be realized by increasing S. But a coplanar line is less compact since the ground return must be sufficiently wide (say 3-4 times the skin depth).

A grounded co-planar waveguide (GCPW) structure is shown in Fig. 5.22. At first glance, the GCPW may seem to have superior characteristics over a normal coplanar line since it does not consume any extra area while the extra ground shields the transmission line from the lossy Si

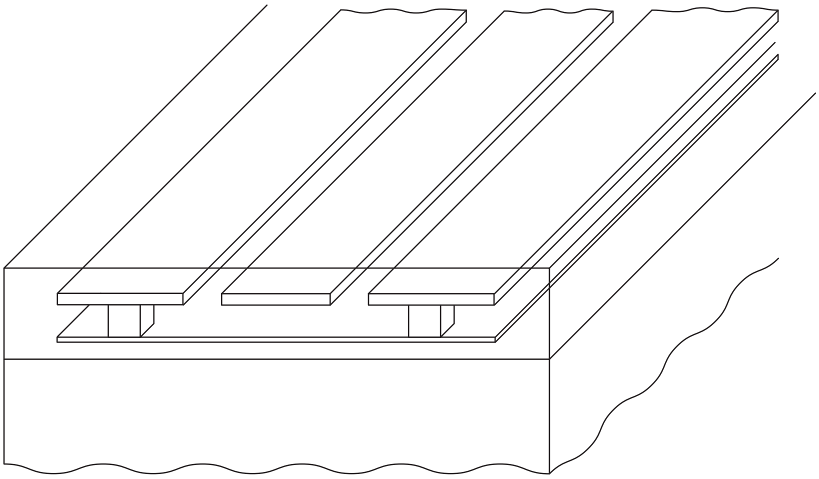


Figure 5.22: A grounded co-planar waveguide (GCPW) structure implemented on in a typical IC process utilizes an array of vias to connect the coplanar ground with the lower ground plane layer.

substrate. To make the comparison more fair, though, we should first note that the ground shield can lower the characteristic impedance, especially if the spacing between the signal and ground plane is smaller than the gap spacing. For instance, as a resonator the GCPW has a lower α and consequently a higher quality factor. At resonance, a shorted quarter wave line appears as a resistor of value $R_{eq} = Z_0/\alpha\ell$. For low power operation, we would like to maximize this resistance and so the ratio Z_0/α is important. In practice one can optimize the GCPW dimensions to realize a higher overall quality factor and a higher equivalent resistance.

The other important consideration is that we should carefully distinguish between the line loss per unit length, α, and the loss that actually matters for a particular application. For example, in many situations, we employ transmission in matching networks as open or short stubs of length $\ell < \lambda/4$. In such a situation, imagine using the transmission line as an equivalent inductor. Then we really care about the series resistance of the structure. Will a grounded co-planar have lower resistance? Certainly this is true at low frequency as adding resistors in parallel should lower the overall resistance. What about high frequency?

Consider a typical CMOS process with a thin metal 1 of thickness $0.2\,\mu$ and a thicker top metal of thickness $1\,\mu$ situated about $6\,\mu$ above from the substrate. Let's compare a short $100\,\mu$ CPW line with gap spacing of $10\,\mu$ to a GCPW line, with the extra ground plane realized on the first metal layer. Since current always flows along the path of least impedance, at low frequency we would expect the ground current to flow on both the top and bottom layer, resulting in a lower overall resistance. But at high frequency, we note that the ground plane path has lower inductance since the spacing between the top and bottom layer is only about $5\,\mu$, as opposed to the gap spacing of $10\,\mu$. In fact, at high frequencies $\omega L' \gg R'$ and so the current flow path is determined by inductive considerations, even if the ground plane path is more lossy. This is certainly conceivable in the case at hand since the ground plane is thinner and has higher sheet resistance. A full-wave simulation confirms the above analysis and the results are summarized in Fig. 5.23. As evident from the figure, at lower frequencies the GCPW has substantially lower normalized resistance R/Z_0 , but at even moderately high frequencies, the situation reverses and the CPW line has much lower normalized resistance. At 20 GHz, for instance, the GCPW line has nearly twice the resistance.

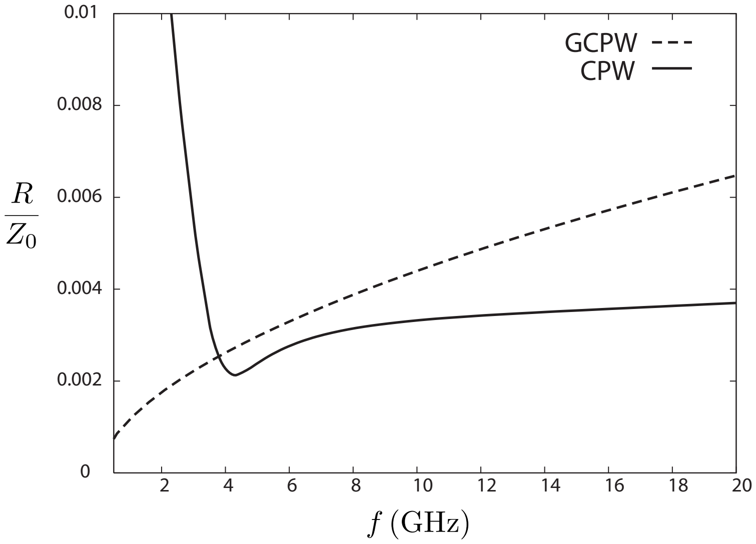


Figure 5.23: A comparison of the resistance of a coplanar waveguide (CPW) structure and a grounded coplanar waveguide (GCPW) structure.

5.7 Transmission Line Circuits

Thus far we have explored the properties of a generic transmission line. We shall now exploit the transmission line as a circuit element, realizing any desired reactance, using it for impedance matching, or as a resonator. The Smith Chart will be introduced as a powerful graphical aid in the design of transmission line circuits. The Smith Chart, a graphical representation of the complex bilinear transform, is also an excellent visualization tool that can give one intuition and insight into transmission line behavior.

5.7.1 Open and Short Transmission Lines

Open Transmission Line

The open line is the dual of the shorted line analyzed in the last chapter. The open transmission line has infinite VSWR and $\rho_L = 1$. At any given point along the transmission line

$$v(z) = V^+(e^{-j\beta z} + e^{j\beta z}) = 2V^+\cos(\beta z) \tag{5.111}$$

whereas the current is given by

$$i(z) = \frac{V^+}{Z_0}(e^{-j\beta z} - e^{j\beta z}) \tag{5.112}$$

or

$$i(z) = \frac{-2jV^+}{Z_0}\sin(\beta z) \tag{5.113}$$

The impedance at any point along the line takes on a simple form

$$Z_{in}(-\ell) = \frac{v(-\ell)}{i(-\ell)} = -jZ_0\cot(\beta\ell) \tag{5.114}$$

This is a special case of the more general transmission line equation with $Z_L = \infty\Omega$. Note that the impedance is purely imaginary since an open lossless transmission line cannot dissipate any power. We have learned, though, that the line stores reactive energy in a distributed fashion.

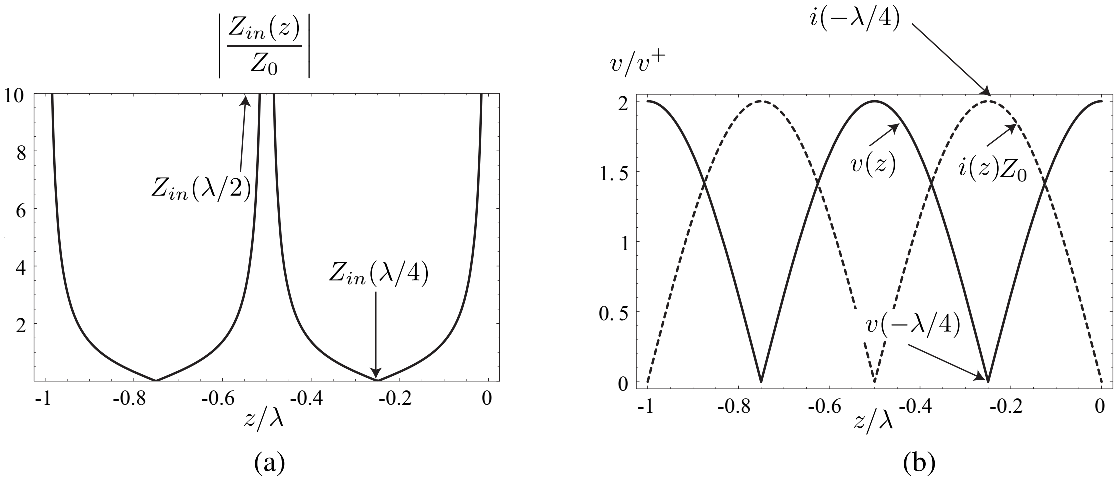


Figure 5.24: (a) The input impedance of an open transmission line as a function of the line length. (b) The voltage and current waveforms on an open line as a function of position.

A plot of the input impedance as a function of z is shown in Fig. 5.24a. As evident from the plot of the current and voltage, shown in Fig. 5.24b, the current must be zero at the end of the line due to the open boundary condition, but it reaches a maximum value at every odd multiple of $\lambda/4$.

The cotangent function takes on zero values when $\beta\ell$ approaches $\pi/2$ modulo 2π. Open transmission line can have zero input impedance! This is particularly surprising since the open load is in effect transformed from an open. A plot of the voltage/current as a function of z is shown below

Notice that if $\ell \ll \lambda/4$, the open line behaves like a lumped capacitor. This is true for any open line at very low frequencies. But even as the frequency increases, and the line approaches a quarter wavelength, $\ell < \lambda/4$, the open line appears as a distributed capacitive reactance. At exactly a quarter wavelength, though, $\ell = \lambda/4$, the line becomes a short circuit. At wavelengths slightly displaced from quarter wavelength, we shall see that the line behaves like a series resonant *LC* circuit. For $\ell > \lambda/4$ but $\ell < \lambda/2$, the line has an inductive reactance. And since the impedance is periodic in $\lambda/2$, the process repeats. This is illustrated in Fig. 5.25.

Shorted Transmission Line

Last chapter we showed that the behavior of the shorted transmission line varies with length, in a completely dual fashion to the open line. If $\ell \ll \lambda/4$, then the line behaves as a lumped inductor. As long as $\ell < \lambda/4$, the reactance is purely inductive. At a quarter wavelength, $\ell = \lambda/4$, the line looks like an open circuit. We shall see that in fact it acts like a resonant parallel *LC* circuit about small deviations from this point. Beyond quarter wavelength, $\ell > \lambda/4$ but $\ell < \lambda/2$, the line becomes capacitive. As before, the process repeats periodically with period $\lambda/2$. This is summarized in Fig. 5.26.

5.7.2 Half-Wave Line

As we have seen, the impedance on a transmission line is periodic about $\lambda/2$. Thus, a section of transmission line of length $\lambda/2$ has an interesting property. Plug into the general T-line equation for any multiple of $\lambda/2$

$$Z_{in}(-m\lambda/2) = Z_0 \frac{Z_L + jZ_0 \tan(-\beta\lambda/2)}{Z_0 + jZ_L \tan(-\beta\lambda/2)} \tag{5.115}$$

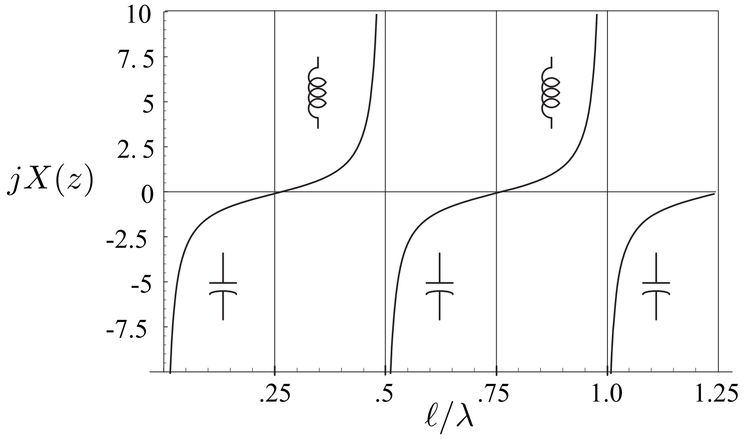


Figure 5.25: The reactance of an open transmission line alternates between capacitive and inductive behavior.

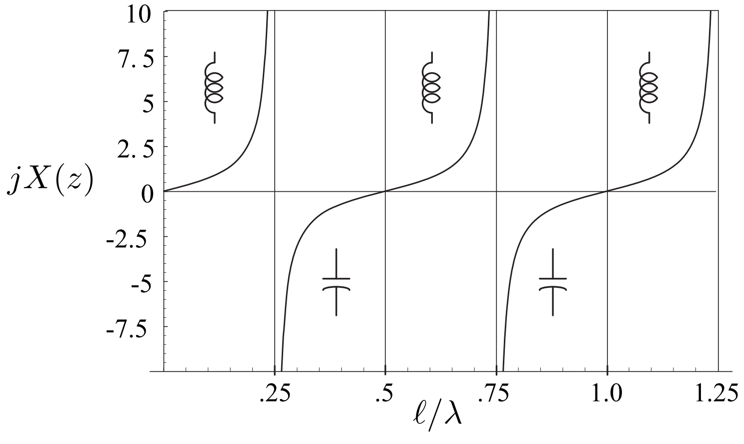


Figure 5.26: The reactance of a shorted transmission line alternates between inductive and capacitive behavior.

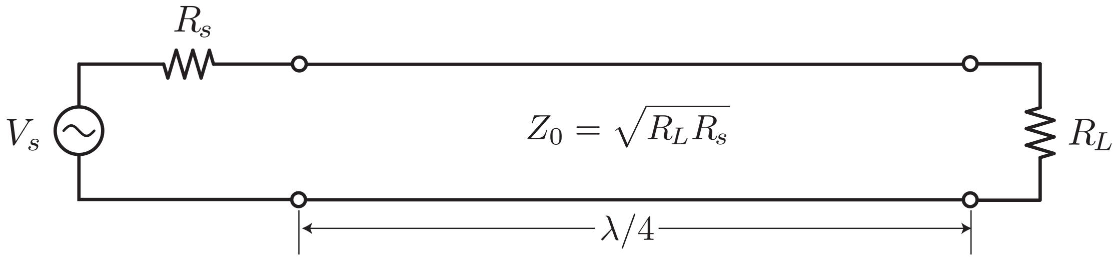


Figure 5.27: The quarter wave transmission line can transform the load resistance to the source resistance by choosing $Z_0 = \sqrt{R_L R_S}$.

Since $\beta\lambda m/2 = \frac{2\pi}{\lambda}\frac{\lambda m}{2} = \pi m$, then $\tan m\pi = 0$ if $m \in \mathscr{Z}$. Or $Z_{in}(-\lambda m/2) = Z_0 \frac{Z_L}{Z_0} = Z_L$. Therefore, as expected, the load impedance does not change as a result of the transmission line. It's as if there were no transmission line connecting the load to the source.

5.7.3 Quarter-Wave Line

Unlike a half-wave line, the quarter-wave line has the most dramatic impact on the load. Starting from the general T-line equation, we see that $\beta\lambda m/4 = \frac{2\pi}{\lambda}\frac{\lambda m}{4} = \frac{\pi}{2}m$, and $\tan m\frac{\pi}{2} = \infty$ if m is an odd integer. We finally have that

$$Z_{in}(-\lambda m/4) = \frac{Z_0^2}{Z_L} \tag{5.116}$$

The $\lambda/4$ line transforms or "inverts" the impedance of the load. This is precisely the observed behavior of the open and short lines. For instance, the shorted line acts like an open when we drive it from a transmission line of length $\lambda/4$, since the zero voltage at the load is transformed to a maximum value at the source. Likewise, the finite current at the load is transformed into zero current at the source.

The property of the $\lambda/4$ line can be exploited to perform impedance matching. As shown in Fig. 5.27, a load resistor R_L is matched to the source resistance R_S by a judicious choice of the line characteristic impedance. In this case, therefore, we equate this to the desired source impedance $Z_{in} = \frac{Z_0^2}{R_L} = R_s$. The quarter-wave line should therefore have a characteristic impedance that is the geometric mean $Z_0 = \sqrt{R_s R_L}$. This only works if the source and load are real resistors.

Since $Z_0 \neq R_L$, the line has a non-zero reflection coefficient

$$SWR = \frac{R_L - \sqrt{R_L R_s}}{R_L + \sqrt{R_L R_s}} \tag{5.117}$$

It also therefore has standing waves on the T-line. The non-unity SWR is given by $\frac{1+|\rho_L|}{1-|\rho_L|}$.

There is a simple explanation for how the $\lambda/4$ line does its magic. Consider a generic lossless transformer ($R_L > R_s$). Thus to make the load look smaller to match to the source, the voltage of the source should be increased in magnitude. But since the transformer is lossless, the current will likewise decrease in magnitude by the same factor. With the $\lambda/4$ transformer, the location of the voltage minimum to maximum is $\lambda/4$ from load (since the load is real). The voltage/current is thus increased/decreased by a factor of $1 + |\rho_L|$ at the load. Hence the impedance decreased by a factor of $(1 + |\rho_L|)^2$

We can evaluate the insertion loss IL of a lossy quarter wave transformer by noting that the power injected into a low-loss transmission line is given by Eq. 5.79

$$P_{in} = \frac{|V^+|^2}{2Z_0}(e^{2\alpha\ell} - |\rho(\lambda/4)|^2 e^{-2\alpha\ell}) \tag{5.118}$$

where

$$|\rho(\lambda/4)| = |\rho_L| e^{-2\alpha\lambda/4} \tag{5.119}$$

The power delivered to the load is simply given by

$$P_L = \frac{|V^+|^2}{2Z_0}(1 - |\rho_L|^2) \tag{5.120}$$

So the insertion loss is given by

$$IL = \frac{P_L}{P_{in}} = \frac{1 - |\rho_L|^2}{e^{2\alpha\lambda/4} - |\rho(\lambda/4)| e^{-2\alpha\lambda/4}} \tag{5.121}$$

The above expression can be simplified. Let the matching ratio be defined as the ratio of the higher to lower resistance

$$m = \frac{R_{hi}}{R_{lo}} \geq 1 \tag{5.122}$$

For instance, if $R_S > R_L$, then $m = R_S/R_L$. Then we can write IL as

$$IL = \frac{1}{\cosh(2\alpha\lambda/4) + \frac{1+m}{2\sqrt{m}} \sinh(2\alpha\lambda/4)} \tag{5.123}$$

In terms of the transmission line $Q = \frac{\beta}{2\alpha}$

$$2\alpha\lambda/4 = \frac{\alpha\lambda}{2} = \frac{\beta}{2Q}\lambda 2 = \frac{\pi}{2Q} \tag{5.124}$$

and thus we can parameterize the insertion loss in terms of the unitless parameters Q and m

$$IL(Q,m) = \frac{1}{\cosh(\frac{\pi}{2Q}) + \frac{1+m}{2\sqrt{m}} \sinh(\frac{\pi}{2Q})} \tag{5.125}$$

For a low loss line $2\alpha\ell \ll 1$ and the above reduces to

$$IL_{ll}(Q,m) = 1 - \frac{(1+m)\pi}{4\sqrt{m}Q} \tag{5.126}$$

A plot of the insertion loss as a function of the matching ratio is shown in Fig. 5.28. Similar to the lumped matching networks, the loss is a function of the matching ratio m, quickly dropping for high ratios.

As already noted, since the reflection coefficient is zero only at a single frequency when the structure is electrically exactly $\lambda/4$ long, the quarter wave transmission line is a narrowband matching network. A plot of the reflection coefficient ρ versus frequency is shown in Fig. 5.29. Note that the higher the matching ratio m, the smaller the bandwidth. In practice, we can usually tolerate a reflection coefficient no larger than a certain amount, $\rho < \rho_m$. For instance, we may tolerate a reflection as large as -10dB. Thus we can calculate the effective bandwidth of the match. For large matching ratio m, the bandwidth can be extended by employing a cascade of quarter wave sections [**Collin2**].

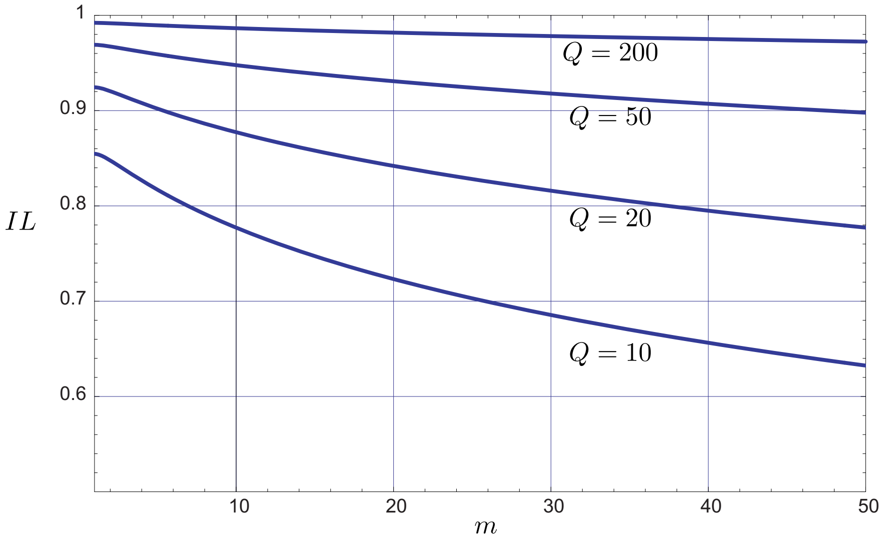


Figure 5.28: The insertion loss of a quarter wave matching network as a function of the matching ratio *m* and the transmission line quality factor *Q*.

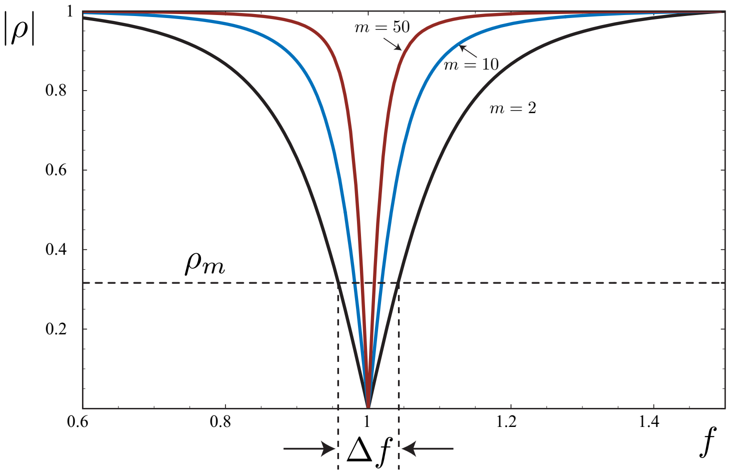


Figure 5.29: The reflection coefficient of a quarterwave matching network as a function of frequency and matching ratio *m*.

5.7.4 Transmission Line Resonance

We would like to now demonstrate that transmission lines acts like resonant circuits around frequencies where the reactance changes sign. Intuitively it seems like a shorted transmission acts like a parallel LC resonant circuit about odd multiples of $\lambda/4$ and like a series LC resonant circuit about even multiple s of $\lambda/4$.

The impedance of a series resonator near resonance $Z(\omega) = j\omega L + \frac{1}{j\omega C} + R$ can be written in terms of the Q factor, $Q = \omega_0 L/R$.

For a small frequency shift from resonance $\delta\omega \ll \omega_0$

$$Z(\omega_0 + \delta\omega) = j\omega_0 L + j\delta\omega L + \frac{1}{j\omega_0 C}\left(\frac{1}{1+\frac{\delta\omega}{\omega_0}}\right) + R \tag{5.127}$$

which can be simplified using the fact that $\omega_0 L = \frac{1}{\omega_0 C}$

$$Z(\omega_0 + \delta\omega) = j2\delta\omega L + R \tag{5.128}$$

Using the definition of Q

$$Z(\omega_0 + \delta\omega) = R\left(1 + j2Q\frac{\delta\omega}{\omega_0}\right) \tag{5.129}$$

For a parallel line, the same formula applies to the admittance

$$Y(\omega_0 + \delta\omega) = G\left(1 + j2Q\frac{\delta\omega}{\omega_0}\right) \tag{5.130}$$

where $Q = \omega_0 C/G$.

Shorted Half-Wave Line Resonance

A shorted transmission line of length ℓ has input impedance of $Z_{in} = Z_0 \tanh(\gamma\ell)$. For a low-loss line, Z_0 is almost real. Expanding the tanh term into real and imaginary parts

$$\tanh(\alpha\ell + j\beta\ell) = \frac{\sinh(2\alpha\ell)}{\cos(2\beta\ell) + \cosh(2\alpha\ell)} + \frac{j\sin(2\beta\ell)}{\cos(2\beta\ell) + \cosh(2\alpha\ell)} \tag{5.131}$$

Since $\lambda_0 f_0 = c$ and $\ell = \lambda_0/2$ (near the resonant frequency), we have $\beta\ell = 2\pi\ell/\lambda = 2\pi\ell f/c = \pi + 2\pi\delta f\ell/c = \pi + \pi\delta\omega/\omega_0$. If the lines are low loss, then $\alpha\ell \ll 1$.

Simplifying the above relation we come to

$$Z_{in} = Z_0\left(\alpha\ell + j\frac{\pi\delta\omega}{\omega_0}\right) \tag{5.132}$$

The above form for the input impedance of the series resonant T-line has the same form as that of the series LRC circuit. We can define equivalent elements

$$R_{eq} = Z_0\alpha\ell = Z_0\alpha\lambda/2 \tag{5.133}$$

$$L_{eq} = \frac{\pi Z_0}{2\omega_0} \tag{5.134}$$

$$C_{eq} = \frac{2}{Z_0\pi\omega_0} \tag{5.135}$$

The equivalent Q factor is given by

$$Q = \frac{1}{\omega_0 R_{eq} C_{eq}} = \frac{\pi}{\alpha\lambda_0} = \frac{\beta_0}{2\alpha} \tag{5.136}$$

For a low-loss line, this Q factor can be made very large. A good T-line might have a Q of 1000 or 10,000 or more. It's difficult to build a lumped circuit resonator with such a high Q factor

Shorted Quarter-Wave Line Resonance

The shorted quarter-wave line will likewise behave like a parallel resonant circuit. For a short-circuited $\lambda/4$ line

$$Z_{in} = Z_0 \tanh(\alpha + j\beta)\ell = Z_0 \frac{\tanh \alpha\ell + j\tan\beta\ell}{1 + j\tan\beta\ell \tanh\alpha\ell} \tag{5.137}$$

Multiply numerator and denominator by $-j\cot\beta\ell$

$$Z_{in} = Z_0 \frac{1 - j\tanh\alpha\ell\cot\beta\ell}{\tanh\alpha\ell - j\cot\beta\ell} \tag{5.138}$$

For $\ell = \lambda/4$ at $\omega = \omega_0$ and $\omega = \omega_0 + \delta\omega$

$$\beta\ell = \frac{\omega_0\ell}{v} + \frac{\delta\omega\ell}{v} = \frac{\pi}{2} + \frac{\pi\delta\omega}{2\omega_0} \tag{5.139}$$

So $\cot\beta\ell = -tan\frac{\pi\delta\omega}{2\omega_0} \approx \frac{-\pi\delta\omega}{2\omega_0}$ and $\tanh\alpha\ell \approx \alpha\ell$, which leads to

$$Z_{in} = Z_0 \frac{1 + j\alpha\ell\pi\delta\omega/2\omega_0}{\alpha\ell + j\pi\delta\omega/2\omega_0} \approx \frac{Z_0}{\alpha\ell + j\pi\delta\omega/2\omega_0} \tag{5.140}$$

This has the same form for a parallel resonant *RLC* circuit

$$Z_{in} = \frac{1}{1/R + 2j\delta\omega C} \tag{5.141}$$

The equivalent circuit elements are

$$R_{eq} = \frac{Z_0}{\alpha\ell} \tag{5.142}$$

$$C_{eq} = \frac{\pi}{4\omega_0 Z_0} \tag{5.143}$$

$$L_{eq} = \frac{1}{\omega_0^2 C_{eq}} \tag{5.144}$$

The quality factor is thus

$$Q = \omega_0 RC = \frac{\pi}{4\alpha\ell} = \frac{\beta}{2\alpha} \tag{5.145}$$

The same as before.

Feynman's Can

In Richard Feynman's class lecture series, he provokes a thought experiment that transforms a simple *LC* tank into a cylindrical resonator. The idea is as follows. Let's say we want to design a high frequency resonator with an inductor and a capacitor. Let's start with a simple coil and capacitor, shown in Fig. 5.30a. How do we increase the resonant frequency? Simply decrease *C* and *L* as much as possible. We can make a small capacitor from two parallel plates by moving the plates far apart. Likewise, we can create a low inductance in shunt with the plates by simply connecting a straight wire from the top plate to the bottom plate. This arrangement is shown in Fig. 5.30b. There is a tradeoff between the distance between the plates and the inductance, since moving the plates further apart will decrease *C* but increase *L*. But we can decrease *L* by putting multiple inductors in parallel, as shown in Fig. 5.30c. The best we can do, though, is to fill the entire surface between the top and bottom plate with wires, forming a cylinder. Thus we see that a "can" is a high frequency resonator!

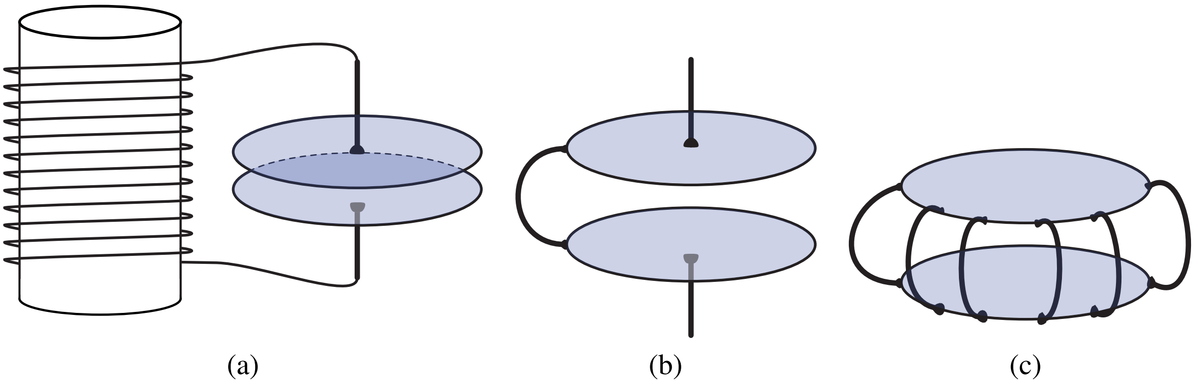


Figure 5.30: (a) A lumped LC resonator is formed by a parallel plate capacitor and a coil inductor. (b) A higher frequency lumped LC resonator employing a single short wire as L. (c) Employing more short wires in parallel further reduces the overall inductance.

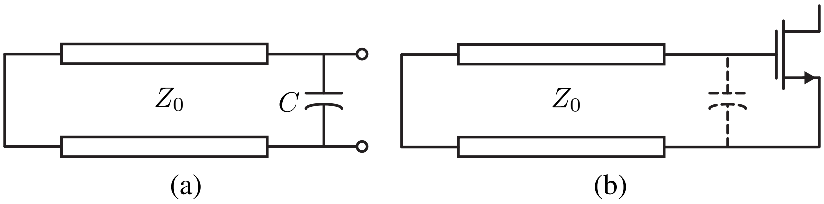


Figure 5.31: (a) A lumped/distributed resonant circuit. The capacitor is in resonance with the transmission line. (b) A common example is the use of a shorted transmission line stub to resonate the input capacitance of a transistor.

Lumped/Distributed Resonant Networks

Often transmission lines are used as resonant elements along with lumped elements. A good example, shown in Fig. 5.31, is a short section of transmission line resonating with the input capacitance of a transistor. For simplicity assume that the lumped input capacitance is lossless. What's the the Q factor of the resulting resonant circuit?

It's important to note that $Q \neq \frac{1}{2}\beta/\alpha$ since this only applies to the transmission line in resonance, when the magnetic and electric energy are equal on the transmission line. In our case, we would like to use the transmission line as an inductor, so we will be concerned with the net magnetic energy on the line. The Q factor is therefore given by

$$Q = 2\omega_0 \frac{\text{net energy stored}}{\text{avg. power loss}} = \frac{2\omega_0(W_m - W_e)}{P_R + P_G} \tag{5.146}$$

where W_m and W_e are the average magnetic and electric energy stored, and P_R represent the "series" resistive losses and P_G the "shunt" conductive losses. Defining the series inductive and shunt capacitive Q we have [**Doan**]

$$Q_L = 2\omega_0 \frac{W_m}{P_R} \tag{5.147}$$

$$Q_C = 2\omega_0 \frac{W_e}{P_G} \tag{5.148}$$

we can express the overall Q as

$$\frac{1}{Q} = \frac{1}{\eta_L Q_L} + \frac{1}{\eta_C Q_C} \tag{5.149}$$

where

$$\eta_L = 1 - \frac{W_e}{W_m} \tag{5.150}$$

and

$$\eta_C = \frac{W_m}{W_e} - 1 \tag{5.151}$$

For a shorted transmission line, under the assumption of low loss, one can show that

$$W_m \approx \frac{1}{2}\frac{LV^{+2}\ell}{Z_0^2}\left(1 + \operatorname{sinc}\left(\frac{4\pi\ell}{\lambda}\right)\right) \tag{5.152}$$

and

$$W_e \approx \frac{1}{2}CV^{+2}\ell\left(1 - \operatorname{sinc}\left(\frac{4\pi\ell}{\lambda}\right)\right) \tag{5.153}$$

Thus we have

$$\frac{1}{\eta_L} = \frac{1}{2\operatorname{sinc}(\frac{4\pi\ell}{\lambda})} + \frac{1}{2} \tag{5.154}$$

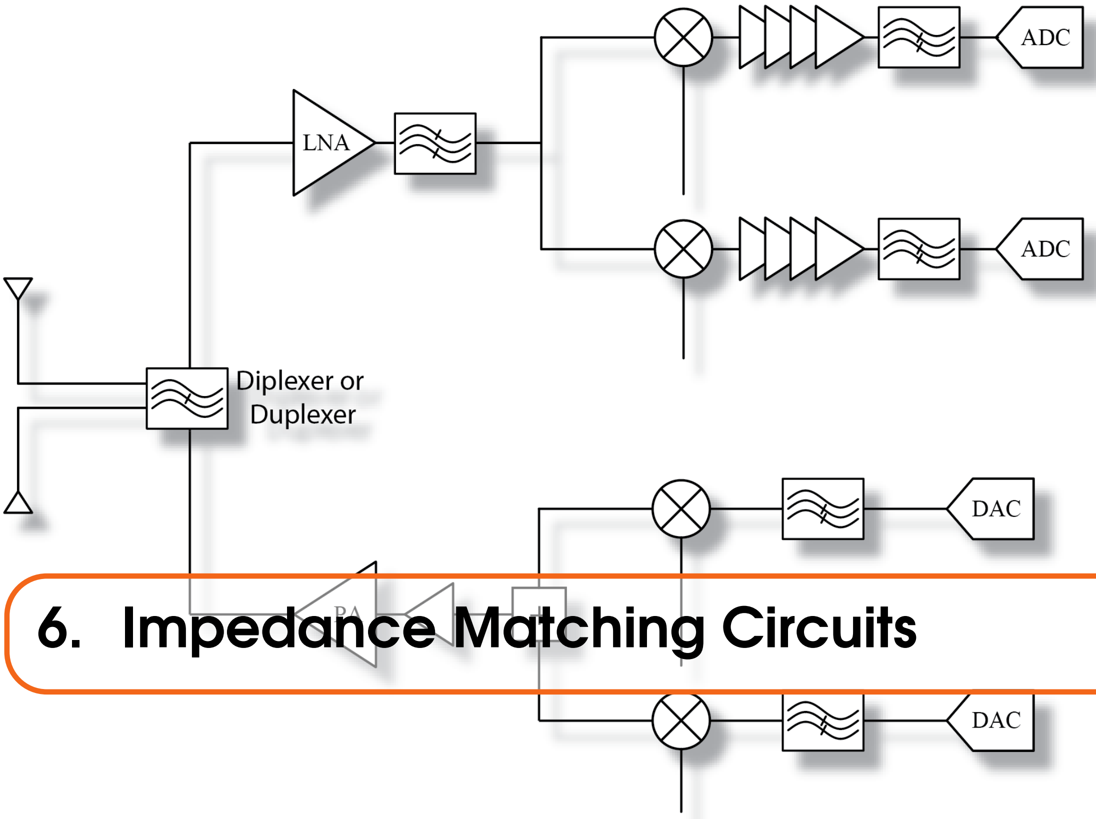
and

$$\frac{1}{\eta_C} = \frac{1}{2\operatorname{sinc}(\frac{4\pi\ell}{\lambda})} - \frac{1}{2} \tag{5.155}$$

For a shorted line, say $\ell \ll \lambda$, then $\eta_C \gg \eta_L$. For instance, if $\ell < 0.1\lambda$, then $\eta_C > 7\eta_L$. The net Q of such a resonant circuit is therefore $Q \approx \eta_L Q_L$.

5.8 References

I have taught this material as part of an undergraduate electromagnetics course and as part of a graduate microwave circuits book, and much of the material is adapted from my lecture notes. In preparing my notes, I have relied on several sources, including Collin's great books [**Collin**] [**Collin2**], Pozar's *Microwave Engineering* book [**Pozar**], and undergraduate electromagnetics books such by Cheng [**Cheng**] and by Inan and Inan [**Inan**].



6. Impedance Matching Circuits

6.1 Introduction

In the words of one experienced designer, “RF design is all about impedance matching.” In this chapter we’d like to show how inductors and capacitors are handy elements at impedance matching. At higher frequencies, we may use transmission lines either as substitutes for lumped components or to modify the impedance of the load by rotating around an “SWR circle”, a term which we will explain in this chapter.

When viewed as a black-box shown in Fig. 6.1, an impedance matcher changes a given load resistance R_L to a source resistance R_S. Without loss of generality, assume $R_S > R_L$, and a power match factor of $m = R_S/R_L$ is desired. In fact any matching network that boosts the resistance by some factor can be flipped over to do the opposite matching.

Since $R_L = v_o/i_o$ and $R_S = v_i/i_i$, we can see that this transformation can be achieved by a voltage gain, $v_i = kv_o$. Assuming the black box is realized with passive elements without memory, power conservation implies

$$i_i v_i = i_o v_o \tag{6.1}$$

thus the current must drop by the same factor, $i_i = k^{-1} i_o$, resulting in

$$Z_{in} = \frac{v_i}{i_i} = \frac{kv_o}{k^{-1}i_o} = k^2\frac{v_o}{i_o} = k^2 R_L \tag{6.2}$$

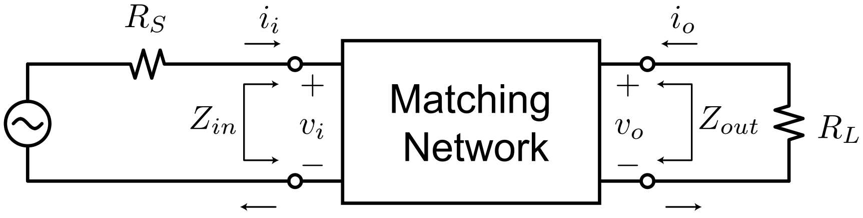


Figure 6.1: A generic matching network as a black box.

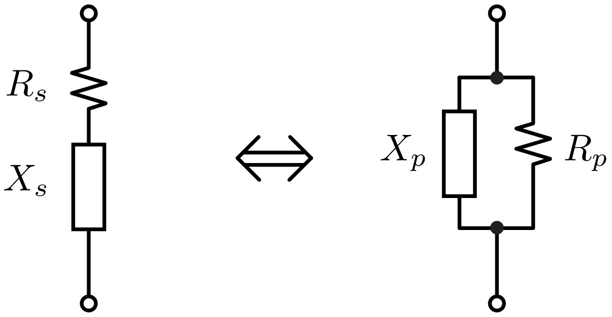


Figure 6.2: A sereis-to-parllel transformation.

which means that $k = \sqrt{m}$ to achieve an impedance match. There are many ways to realize such a circuit block and in this section we'll explore techniques employing inductors and capacitors.

6.1.1 Shunt-Series and Shunt-Series Transformations

The key calculation aid is the series to parallel transformation, which works both ways (parallel to series). Consider the impedance shown in Fig. 6.2, which we wish to represent as a parallel impedance. We can do this at a single frequency as long as the impedance of the series network equals the impedance of the shunt network

$$R_s + jX_s = \frac{1}{\frac{1}{R_p} + \frac{1}{jX_p}} \tag{6.3}$$

Equating the real and imaginary parts

$$R_s = \frac{R_p X_p^2}{R_p^2 + X_p^2} \tag{6.4}$$

$$X_s = \frac{R_p^2 X_p}{R_p^2 + X_p^2} \tag{6.5}$$

which can be simplified by using the definition of Q

$$Q_s = \frac{X_s}{R_s} = \frac{R_p^2 X_p}{R_p X_p^2} = \frac{R_p}{X_p} = Q_p = Q \tag{6.6}$$

Which shows that

$$R_p = R_s(1 + Q^2) \tag{6.7}$$

and

$$X_p = X_s(1 + Q^{-2}) \approx X_s \tag{6.8}$$

where the approximation applies under high Q conditions.

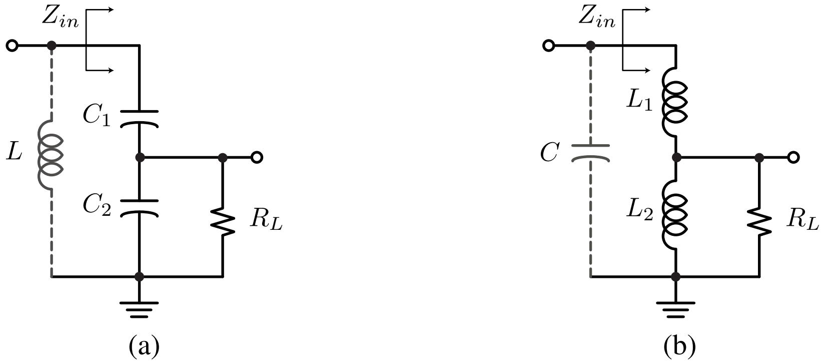


Figure 6.3: (a) A tapped capacitive divider impedance transformer. (b) A tapped inductor impedance transformer. The reactance in the structure can be resonated with an appropriate elements (shown in dashed line).

6.1.2 Capacitive and Inductive Dividers

Perhaps the simplest matching networks are simple voltage dividers. Consider the capacitive voltage divider shown in Fig. 6.3a. At RF frequencies, if $R_L \gg X_2$, then we can see that the circuit will work as advertised. Assuming that negligible current flows into R_L, the current flowing into the capacitors is given by

$$i = \frac{v_i}{j(X_1 + X_2)} \tag{6.9}$$

the voltage across the load is therefore

$$v_o = v_{C_2} = jX_2 \times i = v_i \frac{X_2}{X_1 + X_2} = v_i \frac{1}{1 + \frac{C_2}{C_1}} = kv_i \tag{6.10}$$

which means that the load resistance is boosted by a factor of k^2

$$R_{in} \approx \left(1 + \frac{C_2}{C_1}\right)^2 R_L \tag{6.11}$$

We can arrive at the same destination by using the shunt ↔ series transformation twice. The final value of R_{in} is given by a $1 + Q_2^2$ reduction following by a $1 + Q_s^2$ enhancement

$$R_{in} = \frac{1 + Q_s^2}{1 + Q_2^2} R_L \tag{6.12}$$

where $Q_2 = \frac{R_L}{X_2}$, $X_s = X_1 || X_2'$, and

$$Q_s = \frac{X_s}{R_L}(1 + Q_2^2) \tag{6.13}$$

The final expression is derived after some algebra

$$R_{in} = \frac{R_L}{1 + Q_2^2} + \left(\frac{X_s}{R_L}\right)^2 + \left(\frac{X_s}{X_2}\right)^2 R_L \tag{6.14}$$

Under the assumption that $X_2 \ll R_L$, the final term dominates

$$R_{in} = \left(\frac{X_s}{X_2}\right)^2 R_L \approx \left(1 + \frac{C_2}{C_1}\right)^2 R_L \tag{6.15}$$

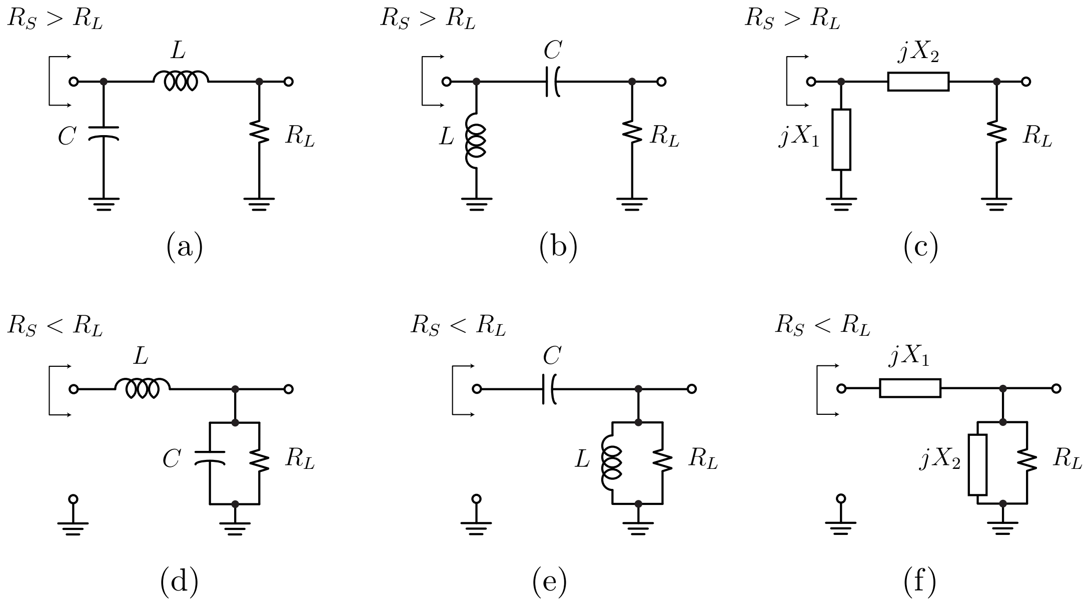


Figure 6.4: Several incarnations of L-matching networks. In (a)-(c) the load is connected in series with the reactance boosting the input resistance. In (d)-(f) the load is in shunt with the reactance, lowering the input resistance.

as expected. The reactance of the capacitive divider can be absorbed by a resonating inductance as shown in Fig. 6.3. In a similar vein, an inductive divider matching circuit can be designed as shown in Fig. 6.3.

6.1.3 An *L*-Match

Consider the L-Matching networks, shown in Fig. 6.4, named due to the topology of the network. We shall see that one direction of the L-match boosts the load impedance (in series with load) whereas the other lowers the load impedance (in shunt with the load). Let's focus on the first two networks shown in Fig. 6.4ab. Here, in absence of the source, we have a simple series *RLC* circuit. Recall that in resonance, the voltage across the reactive elements is Q times larger than the voltage on the load! In essence, that is enough to perform the impedance transformation. Without doing any calculations, you can immediately guess that the impedance seen by the source is about Q^2 larger than R_L. Furthermore, since the circuit is operating in resonance, the net impedance seen by the source is purely real. To be sure, let's do the math.

A quick way to accomplish this feat is to begin with the series to parallel transformation, as shown in Fig. 6.5, where the load resistance in series with the inductor is converted to an equivalent parallel load equal to

$$R_p = (1+Q^2)R_L \tag{6.16}$$

where $Q = X_L/R_L$, and $X_L' = X_L(1+Q^{-2})$. The circuit is now nothing but a parallel *RLC* circuit and it's clear that at resonance the source will see only R_p, or a boosted value of R_L. The boosting factor is indeed equal to Q^2+1, very close to the value we guessed from the outset.

To gain insight into the operation of Fig. 6.4d-f, consider an Norton equivalent of the same circuit shown in Fig. 6.6. Now the circuit is easy to understand since it's simply a parallel resonant circuit. We known that at resonance the current through the reactances is Q times larger than the current in the load. Since the current in the series element (L in Fig. 6.4d) is controlled by the

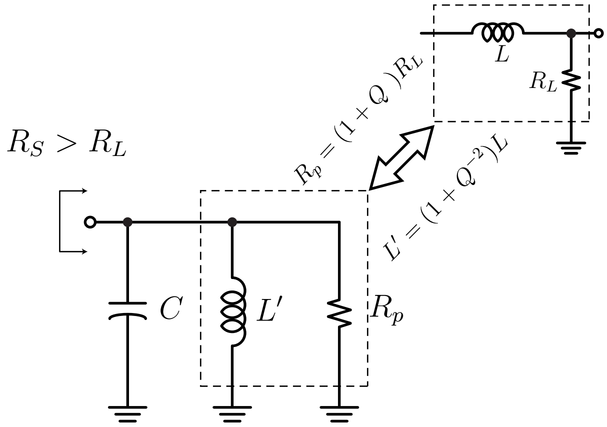


Figure 6.5: The transformed L matching network into a parallel *RLC* equivalent circuit.

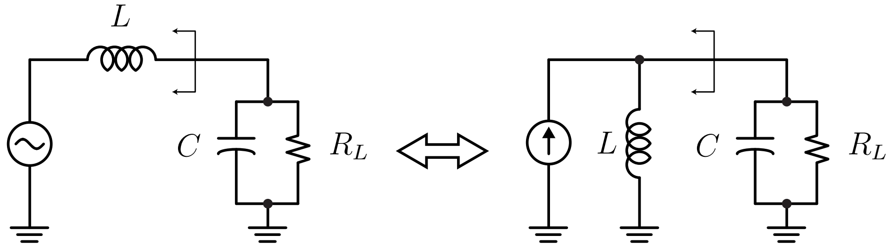


Figure 6.6: The source voltage driving the L matching network can be transformed into an equivalent Norton current source.

source voltage, we can immediately see that $i_s = Qi_L$, thus providing the required current gain to lower the load resistance by a factor of Q^2.

As you may guess, the mathematics will yield a similar result. Simply do a parallel to series transformation of the load to obtain

$$R_s = \frac{R_p}{1+Q^2} \tag{6.17}$$

$$X_p' = \frac{X_p}{1+Q^{-2}} \tag{6.18}$$

The resulting circuit is a simple series *RLC* circuit. At resonance, the source will only see the reduced series resistance R_s.

L-Match Design Equations

The following design procedure applies to an *L*-match using the generic forms of Fig. 6.4c,f. The actual choice between Fig. 6.4a,d and Fig. 6.4b,e depends on the application. For instance Fig. 6.4b,e provide AC coupling (DC isolation) which may be required in many applications. In other applications a common DC voltage may be needed, making the networks of Fig. 6.4a,d the obvious choice.

Let $R_{hi} = \max(R_S, R_L)$ and $R_{lo} = \min(R_S, R_L)$. The L-matching networks shown in Fig. 6.4 are designed as follows:

1. Calculate the boosting factor $m = \frac{R_{hi}}{R_{lo}}$.
2. Compute the required circuit Q by $(1+Q^2) = m$, or $Q = \sqrt{m-1}$.
3. Pick the required reactance from the Q. If you're boosting the resistance, e.g. $R_S > R_L$, then $X_s = Q \cdot R_L$. If you're dropping the resistance, $X_p = \frac{R_L}{Q}$.
4. Compute the effective resonating reactance. If $R_S > R_L$, calculate $X_s' = X_s(1+Q^{-2})$ and set the shunt reactance in order to resonate, $X_p = -X_s'$. If $R_S < R_L$, then calculate $X_p' = \frac{X_p}{1+Q^{-2}}$ and set the series reactance in order to resonate, $X_s = -X_p'$.
5. For a given frequency of operation, pick the value of L and C to satisfy these equations.

Insertion Loss of an *L*-Matching Network

We'd like to include the losses in our passive elements into the design of the matching network. The most detrimental effect of the component Q is the insertion loss which reduces the power transfer from source to load.

Let's begin by using our intuition to derive an approximate expression for the loss. Note that the power delivered to the input of the matching network P_{in} can be divided into two components

$$P_{in} = P_L + P_{diss} \tag{6.19}$$

where P_L is the power delivered to the load and P_{diss} is the power dissipated by the non-ideal inductors and capacitors. The insertion loss is therefore given by

$$IL = \frac{P_L}{P_{in}} = \frac{P_L}{P_L + P_{diss}} = \frac{1}{1 + \frac{P_{diss}}{P_L}} \tag{6.20}$$

Recall that for the equivalent series *RLC* circuit in resonance, the voltages across the reactances are Q times larger than the voltage across R_L. We can show that the reactive power is also a factor of Q larger. For instance the energy in the inductor is given by

$$W_m = \frac{1}{4}Li_s^2 = \frac{1}{4}\frac{v_s^2}{4R_S^2}L \tag{6.21}$$

or

$$\omega_0 \times W_m = \tfrac{1}{4}\frac{v_s^2}{4R_S}\frac{\omega_0 L}{R_S} = \tfrac{1}{2}\frac{v_s^2}{8R_S}Q = \tfrac{1}{2}P_L \times Q \tag{6.22}$$

where P_L is the power to the load at resonance

$$P_L = \frac{v_L^2}{2R_S} = \frac{v_s^2}{4 \cdot 2 \cdot R_S} = \frac{v_s^2}{8R_S} \tag{6.23}$$

The total reactive power is thus exactly Q times larger than the power in the load

$$\omega_0(W_m + W_e) = Q \times P_L \tag{6.24}$$

By the definition of the component Q_c factor, the power dissipated in the non-ideal elements of net quality factor Q_c is simply

$$P_{diss} = \frac{P_L \cdot Q}{Q_c} \tag{6.25}$$

which by using Eq. 6.20 immediately leads to the following expression for the insertion loss

$$IL = \frac{1}{1 + \frac{Q}{Q_c}} \tag{6.26}$$

The above equation is very simple and insightful. Note that using a higher network Q, e.g. a higher matching ratio, incurs more insertion loss with the simple single stage matching network. Furthermore, the absolute component Q is not important but only the component Q_c normalized to the network Q. Thus if a low matching ratio is needed, the actual components can be moderately lossy without incurring too much insertion loss.

Also note that the the actual inductors and capacitors in the circuit can be modeled with very complicated sub-circuits, with several parasitics to model distributed and skin effect, but in the end, at a given frequency, one can calculate the equivalent component Q_c factor and use it in the above equation.

Note that Q_c is the net quality factor of the passive elements. If one element dominates, such as a low-Q inductor, then Q_L can be used in its place. The exact analysis for a lossy inductor and capacitor is simple enough and yields an expression that is identical to Eq. 6.26 when only inductor losses are taken into account but differs when both inductor and capacitors losses are included

$$IL = \frac{1}{1 + \frac{Q}{Q_L}}\frac{1}{1 + \frac{Q}{Q_C}} \tag{6.27}$$

which can be written as

$$IL = \frac{1}{1 + Q\left(Q_L^{-1} + Q_C^{-1}\right) + \frac{Q^2}{Q_L Q_C}} \tag{6.28}$$

which equals the general expression we derived under “high-Q” conditions, e.g. $Q_c \gg Q$.

When completing the design with real elements, it’s also necessary to shift the component values slightly due to the extra loss. While formulas for these perturbations can be calculated, a modern computer and optimizer really make this exercise unnecessary.

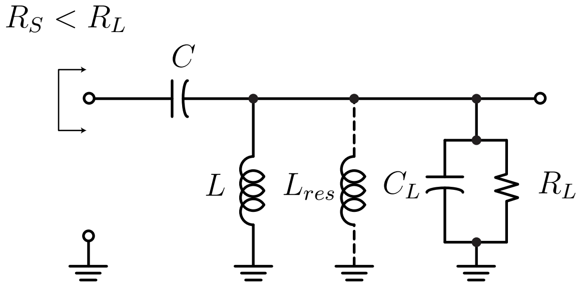


Figure 6.7: The complex load C_L in parallel with R_L is matched to a real source impedance by first applying an parallel inductor L_{res} to resonate out C_L. The load can now be matched using a standard matching network. In the final design, the resonating L_{res} can be simply absorbed into L.

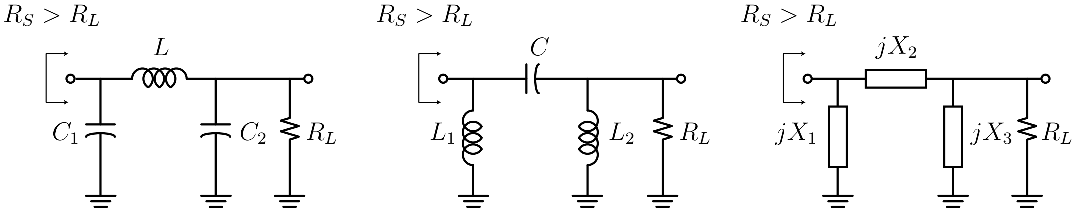


Figure 6.8: Several incarnations of a Π matching network. The first is a low-pass structure, the second a high-pass structure. The third is a general Π network.

Reactance Absorption

In most situations the load and source impedances are often complex and our discussion so far only applies to real load and source impedances. An easy way to handle complex loads is to simply absorb them with reactive elements. For example, consider the complex load shown in Fig. 6.7. To apply an L-matching circuit, we can begin by simply resonating out the load reactance at the desired operating frequency. For instance, we add an inductance L_{res} in shunt with the capacitor to produce a real load. From here the design procedure is identical. Note that we can absorb the inductor L_{res} into the shunt L-matching element.

From now onwards we can simply discuss the real matching problem since a complex load or source can be handled in a similar fashion. Often there are multiple ways to perform the absorption with each choice yielding slightly different network properties such as Q (bandwidth), and different frequency selectivity (e.g. low-pass, high-pass, bandpass).

6.1.4 A Π-Match

The L-Match circuit is simple and elegant but is somewhat constrained. In particular, we cannot freely choose the Q of the circuit since it is fixed by the required matching factor m. This restriction is easily solved with the Π-Matching circuit, also named from its topology, shown in Fig. 6.8. The idea behind the Π match can be easily understood by studying the cascade of two back-to-front L matches as shown in Fig. 6.9. In this circuit the first L match will lower the load impedance to an intermediate value R_i

$$R_i = \frac{R_L}{1+Q_1^2} \tag{6.29}$$

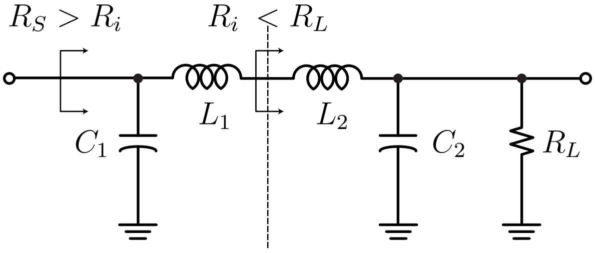


Figure 6.9: The Π network can be decomposed into a back-to-front cascade of two L matching networks. The impedance is first reduce down to $R_i < R_L$, then increased back up to $R_S > R_L > R_i$.

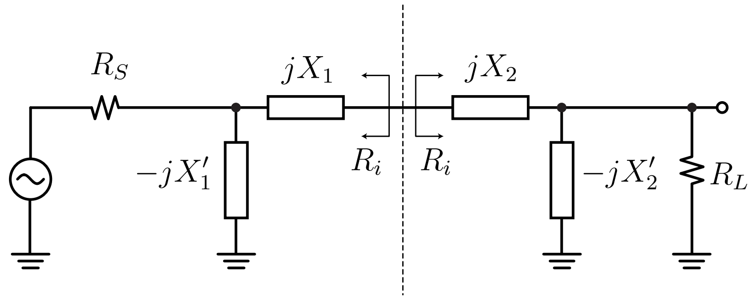


Figure 6.10: The reflected input and output impedance are both equal to R_i at the center of the Π network.

or

$$Q_1 = \sqrt{\frac{R_L}{R_i} - 1} \tag{6.30}$$

Since $R_i < R_L$, the second L match needs to boost the value of R_i up to R_s. The Q of the second L network is thus

$$Q_2 = \sqrt{\frac{R_S}{R_i} - 1} > \sqrt{\frac{R_S}{R_L} - 1} \tag{6.31}$$

When we combine the two L networks, we obtain a Π network with a higher Q than possible with a single stage transformation. In general the Q, or equivalently the bandwidth $B = \frac{\omega_0}{Q}$, is a free

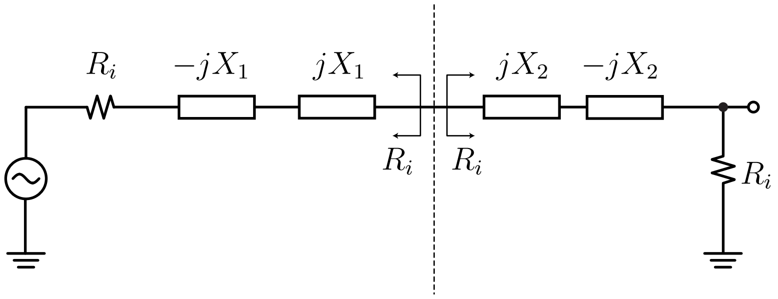


Figure 6.11: The L sections can be converted into series sections to produce one big LCR circuit.

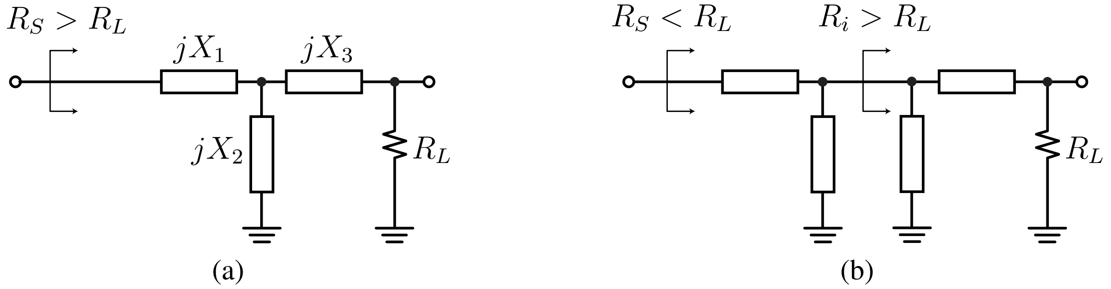


Figure 6.12: (a) The T-matching network can be decomposed into two front-to-back L sections. (b) The first L section boosts the resistance to a value of $R_i > R_L$ and the second L structure drops the impedance to $R_S < R_i$.

parameter that can be chosen at will for a given application. Note that when the source is connected to the input, the circuit is symmetric about the center, as shown in Fig. 6.10. Now it's rather easy to compute the network Q by drawing a series equivalent circuit about the center of the structure, as shown in Fig. 6.11. If the capacitors and inductors in series are combined, the result is a simple *RLC* circuit with Q given by

$$Q = \frac{X_1 + X_2}{2R_i} = \frac{Q_1 + Q_2}{2} \tag{6.32}$$

It's important to note the inclusion of the source resistance when calculating the network Q as we are implicitly assuming a power match. In a power amplifier, the source impedance may be different and the above calculation should take that into consideration. For instance, if the PA is modeled as a high impedance current source (Class A/B operation), then the factor of 2 disappears. The design procedure begins with the specification of the network Q. Eq. 6.32 is then used to find R_i, and from there the L-match procedure outlined above takes over.

6.1.5 A T-Match

The T-matching network, shown in Fig. 6.12a, is the dual of the Π network. By now you can see that the names all correspond the physical topology of the circuit. The T network can also be decomposed into a cascade of two back-to-front L networks, as shown in Fig. 6.12b. The first L transforms the resistance up to some intermediate value $R_i > R_S$, and the second L transforms the resistance back down to R_S. Thus the net Q is higher than a single stage match. The network Q can be derived in an analogous fashion and yields the same solution

$$Q = \tfrac{1}{2}\left(\sqrt{\frac{R_i}{R_L} - 1} + \sqrt{\frac{R_i}{R_S} - 1}\right) \tag{6.33}$$

6.1.6 Multi-Section Low Q Matching

We have seen that the Π and T matching networks are essentially two stage networks which can boost the network Q. In many applications we actually would like to achieve the opposite effect, e.g. low network Q is desirable in broadband applications. Furthermore, a low Q design is less susceptible to process variations. Also, a lower Q network lowers the loss of the network, as evident by examining Eq. 6.26.

To lower the Q of an L matching network, we can employ more than one stage to change the impedance in smaller steps. Recall that $Q = \sqrt{m-1}$, and a large m factor requires a high Q match. If we simply change the impedance by a factor $k < m$, the Q of the first L section is

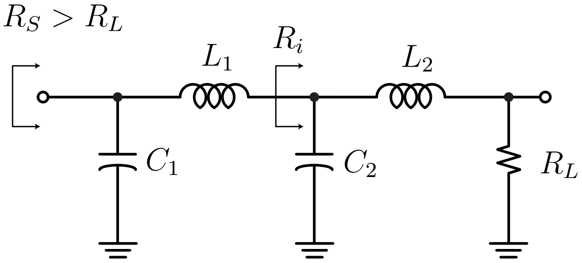


Figure 6.13: A two-stage low-pass L matching network. The first stage steps up the intermediate resistance $R_S < R_i < R_L$, thus lowering the Q over a single stage design.

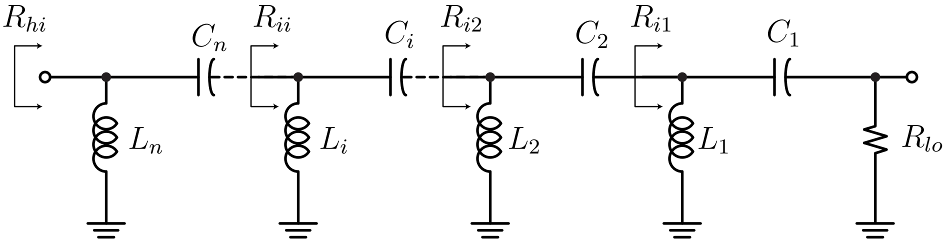


Figure 6.14: A high-pass multi-section L matching network.

reduced. Likewise, a second L section will further change the resistance to the desired R_S with a step size $l < m$, where $l \cdot k = m$. An example two-stage network is shown in Fig. 6.13. Reflecting all impedances to the center of the network, the real part of the impedance looking left or right is R_i at resonance. Thus the power dissipation is equal for both networks. The overall Q is thus given by

$$Q = \frac{\omega(W_{s1} + W_{s2})}{P_{d1} + P_{d2}} = \frac{\omega W_{s1}}{2P_d} + \frac{\omega W_{s2}}{2P_d} = \frac{Q_1 + Q_2}{2} \tag{6.34}$$

$$Q = \tfrac{1}{2}\left(\sqrt{\frac{R_i}{R_L} - 1} + \sqrt{\frac{R_S}{R_i} - 1}\right) \tag{6.35}$$

Note the difference between the above and Eq. 6.33. The R_i term appears once in the denominator and once in the numerator since it's an intermediate value. What's the lowest Q achievable? To find out, take the derivative of Eq.6.35 with respect to R_i and solve for the minimum

$$R_{i,\text{opt}} = \sqrt{R_L R_S} \tag{6.36}$$

which results in a Q approximately lower by a square root factor

$$Q_{opt} = \sqrt{\sqrt{\frac{R_S}{R_L}} - 1} \approx m^{1/4} \tag{6.37}$$

It's clear that the above equations apply to the opposite case when $R_L > R_S$ by simply interchanging the role of the source and the load.

To even achieve a lower Q, we can keep adding sections as shown in Fig. 6.14. The optimally low Q value is obtained when the intermediate impedances are stepped in geometric progression

$$\frac{R_{i1}}{R_{lo}} = \frac{R_{i2}}{R_{i1}} = \frac{R_{i3}}{R_{i2}} = \cdots = \frac{R_{hi}}{R_{in}} = 1 + Q^2 \tag{6.38}$$

where $R_{hi} = \max(R_S, R_L)$ and $R_{lo} = \min(R_S, R_L)$. In the limit that $n \to \infty$, we take very small "baby" steps from R_{lo} to R_{hi} and the circuit starts to look like a tapered transmission line. Multiplying each term in the above equation

$$\frac{R_{i1}}{R_{lo}} \cdot \frac{R_{i2}}{R_{i1}} \cdot \frac{R_{i3}}{R_{i2}} \cdot \cdots \cdot \frac{R_{hi}}{R_{in}} = \frac{R_{hi}}{R_{lo}} = (1+Q^2)^N \tag{6.39}$$

which results in the optimal Q factor for the overall network

$$Q = \sqrt{\left(\frac{R_{hi}}{R_{lo}}\right)^{1/N} - 1} \tag{6.40}$$

The loss in the optimal multi-section line can be calculated as follows. Using the same approach as Sec. 6.1.3, note that the total power dissipated in the matching network is given by

$$P_{diss} = \frac{NQP_L}{Q_u} \tag{6.41}$$

where N sections are used, each with equal Q due to the condition set forth by Eq. 6.38. This leads to the following expression

$$IL = \frac{1}{1 + N\frac{Q}{Q_u}} \tag{6.42}$$

or

$$IL = \frac{1}{1 + \frac{N}{Q_u}\sqrt{\left(\frac{R_{hi}}{R_{lo}}\right)^{1/N} - 1}} \tag{6.43}$$

It's interesting to observe that this expression has an optimum for a particular value of N. It's easy enough to plot IL for a few values of N to determine the optimal number of sections. Intuitively adding sections can decrease the insertion loss since it also lowers the network Q factor. Adding too many sections, though, can counterbalance this benefit.

6.2 The Smith Chart

The Smith Chart, shown in Fig. 6.15, is simply a graphical calculator for computing impedance as a function of reflection coefficient $z = f(\rho)$. More importantly, many problems can be easily visualized with the Smith Chart. This visualization leads to insights about the behavior of transmission lines. All the knowledge is coherently and compactly represented by the Smith Chart[1].

6.2.1 Smith Chart Construction

Let's begin with the voltage on the line

$$v(z) = v^+(z) + v^-(z) = V^+(e^{-\gamma z} + \rho_L e^{\gamma z})$$

Recall that we can define the reflection coefficient anywhere by taking the ratio of the reflected wave to the forward wave

$$\rho(z) = \frac{v^-(z)}{v^+(z)} = \frac{\rho_L e^{\gamma z}}{e^{-\gamma z}} = \rho_L e^{2\gamma z} \tag{6.44}$$

[1] There are purely aesthetic reasons to study the Smith Chart that arise from deep mathematical connections with the complex bilinear transform.

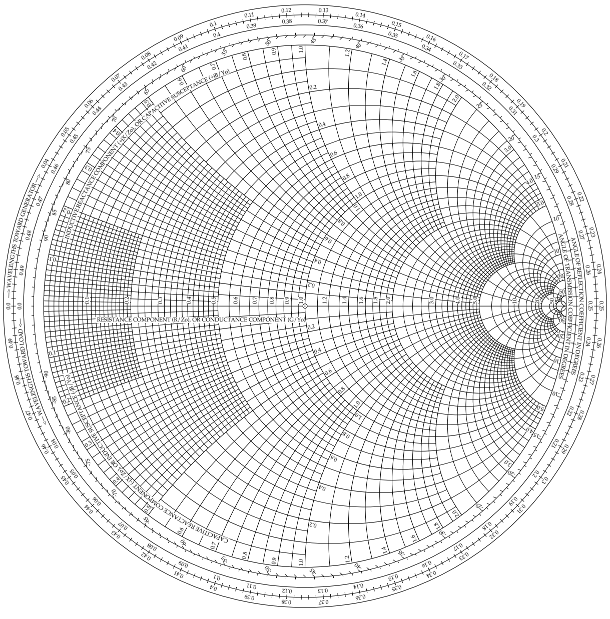


Figure 6.15: The Smith Chart.

Therefore the impedance on the line

$$Z(z) = \frac{v^{+}e^{-\gamma z}(1+\rho_L e^{2\gamma z})}{\frac{v^{+}}{Z_0}e^{-\gamma z}(1-\rho_L e^{2\gamma z})} \tag{6.45}$$

can be expressed in terms of $\rho(z)$

$$Z(z) = Z_0 \frac{1+\rho(z)}{1-\rho(z)} \tag{6.46}$$

It is extremely fruitful to work with normalized impedance values $z = Z/Z_0$

$$z(z) = \frac{Z(z)}{Z_0} = \frac{1+\rho(z)}{1-\rho(z)} \tag{6.47}$$

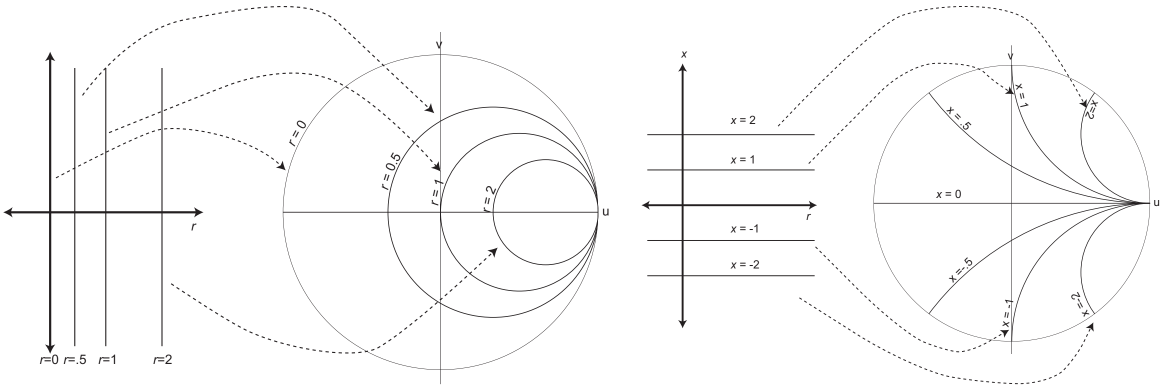
Let the normalized impedance be written as $z = r + jx$ (note small case). The reflection coefficient is "normalized" by default since for passive loads $|\rho| \leq 1$. Let $\rho = u + jv$. Now simply equate the $\Re$ and $\Im$ components in the above equation

$$r + jx = \frac{(1+u)+jv}{(1-u)-jv} = \frac{((1+u+jv)(1-u+jv)}{(1-u)^2+v^2} \tag{6.48}$$

To obtain the relationship between the (r,x) plane and the (u,v) plane

$$r = \frac{1-u^2-v^2}{(1-u)^2+v^2} \tag{6.49}$$

$$x = \frac{v(1-u)+v(1+u)}{(1-u)^2+v^2} \tag{6.50}$$



(a) (b)

Figure 6.16: Graphical illustration of mappings from the (r,x) plane to the complex reflection coefficient unit circuit (u,v).

The above equations can be simplified and put into a nice form. If you remember your grade school algebra, you can derive the following equivalent equations

$$\left(u-\frac{r}{1+r}\right)^2+v^2=\frac{1}{(1+r)^2} \tag{6.51}$$

$$(u-1)^2+\left(v-\frac{1}{x}\right)^2=\frac{1}{x^2} \tag{6.52}$$

These are circles in the (u,v) plane! Circles are good! We see that vertical and horizontal lines in the (r,x) plane (complex impedance plane) are transformed to circles in the (u,v) plane (complex reflection coefficient).

Some special mappings, shown in Fig. 6.16ab, are worth noting

- $r=0$ maps to $u^2+v^2=1$ (unit circle)
- $r=1$ maps to $(u-1/2)^2+v^2=(1/2)^2$ (matched real part)
- $r=.5$ maps to $(u-1/3)^2+v^2=(2/3)^2$ (load R less than Z_0)
- $x=\pm 1$ maps to $(u-1)^2+(v\mp 1)^2=1$
- $x=\pm 2$ maps to $(u-1)^2+(v\mp 1/2)^2=(1/2)^2$
- $x=\pm 1/2$ maps to $(u-1)^2+(v\mp 2)^2=2^2$

Inductive reactance maps to upper half of unit circle. Capacitive reactance maps to lower half of unit circle. A simpler Smith Chart, shown in Fig. 6.17, can be constructed from the above mappings.

6.2.2 Load on Smith Chart

As shown in Fig. 6.18, we can simply plot z_L on the Smith Chart. One can read off ρ_L as a polar complex number. To read off the impedance on the T-line at any point on a lossless line, simply move on a circle of constant radius since $\rho(z)=\rho_L e^{2j\beta}$. Moving towards generator means $\rho(-\ell)=\rho_L e^{-2j\beta\ell}$, or clockwise motion. For a lossy line, this corresponds to a spiral motion. We're back to where we started when $2\beta\ell=2\pi$, or $\ell=\lambda/2$. Thus the impedance is periodic (as we know). Since SWR is a function of $|\rho|$, a circle at origin in (u,v) plane is called an SWR circle.

Since SWR is a function of $|\rho|$, a circle at origin in (u,v) plane is called an SWR circle. Recall the voltage max occurs when the reflected wave is in phase with the forward wave, so $\rho(z_{min})=|\rho_L|$.

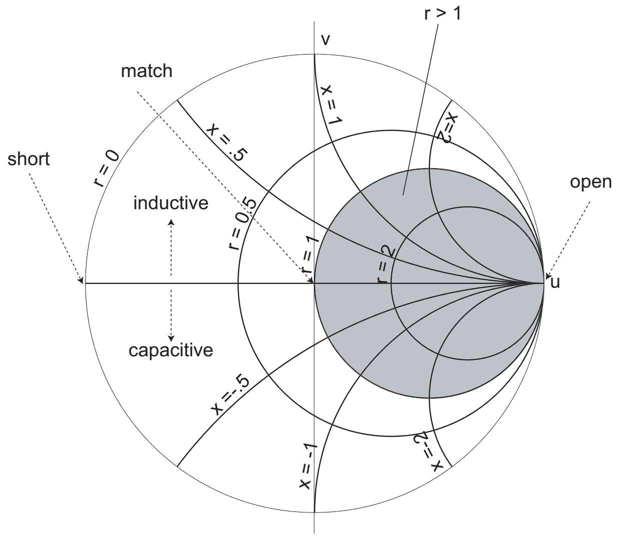
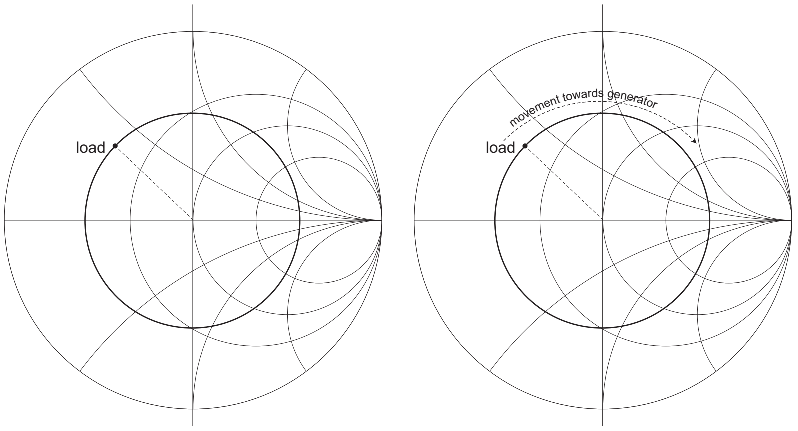


Figure 6.17: A simple Smith Chart illustrating some key points.



(a) (b)

Figure 6.18: (a) A load impedance is plotted on the Smith Chart. (b) By moving clockwise on a circle of constant radius, we can read off the impedance at any point on the transmission line.

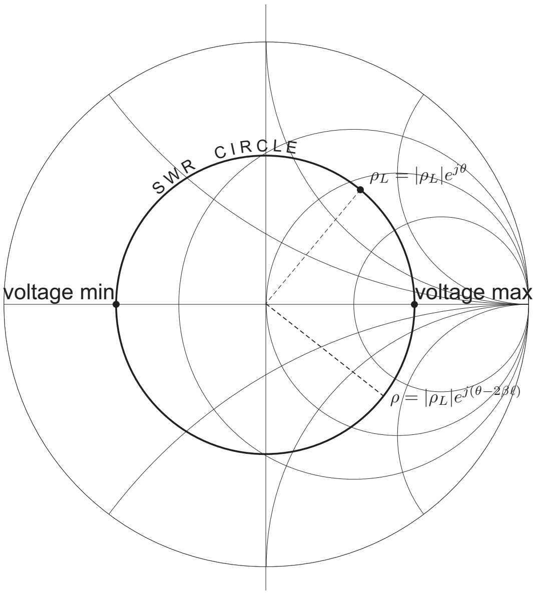

SWR CIRCLE
$\rho_L = |\rho_L|e^{j\theta}$
voltage min
voltage max
$\rho = |\rho_L|e^{j(\theta-2\beta\ell)}$

Figure 6.19: A SWR circle on the Smith Chart.

This corresponds to the intersection of the SWR circle with the positive real axis, shown in Fig. 6.19. Likewise, the intersection with the negative real axis is the location of the voltage minimum.

Example 6: This example will illustrate the utility of the Smith Chart for visualization. Let's prove that if Z_L has an inductance reactance, then the position of the first voltage maximum occurs before the voltage minimum as we move towards the generator. A visual proof is easy using Smith Chart. On the Smith Chart start at any point in the upper half of the unit circle. Moving towards the generator corresponds to clockwise motion on a circle. Therefore we will always cross the positive real axis first and then the negative real axis. This means we must observe a voltage minimum before the maximum. The opposite is true for a capacitive load.

6.2.3 The Admittance Chart

Since $y = 1/z = \frac{1-\rho}{1+\rho}$, you can imagine that an Admittance Smith Chart looks very similar. In fact everything is switched around a bit and you can buy or construct a combined admittance/impedance smith chart. You can also use an impedance chart for admittance if you simply map $x \to b$ and $r \to g$. Be careful as the capacitors are now on the top of the chart and the inductors on the bottom. The short and open likewise swap positions.

Sometimes you may need to work with both impedances and admittances. This is easy on the Smith Chart due to the impedance inversion property of a $\lambda/4$ line

$$Z' = \frac{Z_0^2}{Z} \tag{6.53}$$

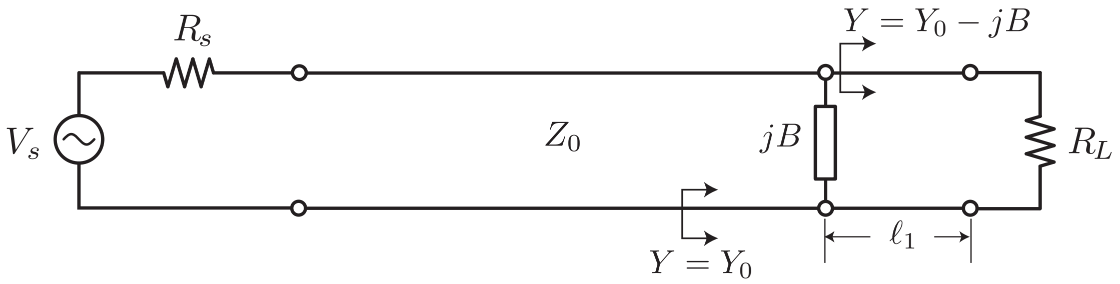


Figure 6.20: A section of transmission line and a shunt reactance can be used to match a load to a desired impedance .

If we normalize Z' we get y

$$\frac{Z'}{Z_0} = \frac{Z_0}{Z} = \frac{1}{z} = y \tag{6.54}$$

Thus if we simply rotate π degrees on the Smith Chart and read off the impedance, we're actually reading off the admittance! Rotating π degrees is easy. Simply draw a line through origin and z_L and read off the second point of intersection on the SWR circle

6.3 Transmission Line Matching Networks

6.3.1 Matching with Lumped Elements

Recall the input impedance looking into a T-line varies periodically

$$Z_{in}(-\ell) = Z_0 \frac{Z_L + jZ_0 \tan(\beta\ell)}{Z_0 + jZ_L \tan(\beta\ell)} \tag{6.55}$$

As shown in Fig. 6.20, move a distance ℓ_1 away from the load such that the real part of Z_{in} has the desired value. Then place a shunt or series impedance on the T-line to obtain desired reactive part of the input impedance (e.g. zero reactance for a real match). For instance, for a shunt match, the input admittance looking into the line is

$$y(z) = Y(z)/Y_0 = \frac{1 - \rho_L e^{j2\beta z}}{1 + \rho_L e^{j2\beta z}} \tag{6.56}$$

At a distance ℓ_1 we desire the normalized admittance to be $y_1 = 1 - jb$. Substitute $\rho_L = \rho e^{j\theta}$ and solve for ℓ_1 and let $\psi = 2\beta z + \theta$

$$\frac{1 - \rho e^{j\psi}}{1 + \rho e^{j\psi}} = \frac{1 - \rho^2 - j2\rho \sin\psi}{1 + 2\rho \cos\psi + \rho^2} \tag{6.57}$$

Solve for ψ (and then ℓ_1) from $\Re(y) = 1$

$$\psi = \theta - 2\beta\ell = \cos^{-1}(-\rho) \tag{6.58}$$

$$\ell_1 = \frac{\theta - \psi}{2\beta} = \frac{\lambda}{4\pi}\left(\theta - \cos^{-1}(-\rho)\right) \tag{6.59}$$

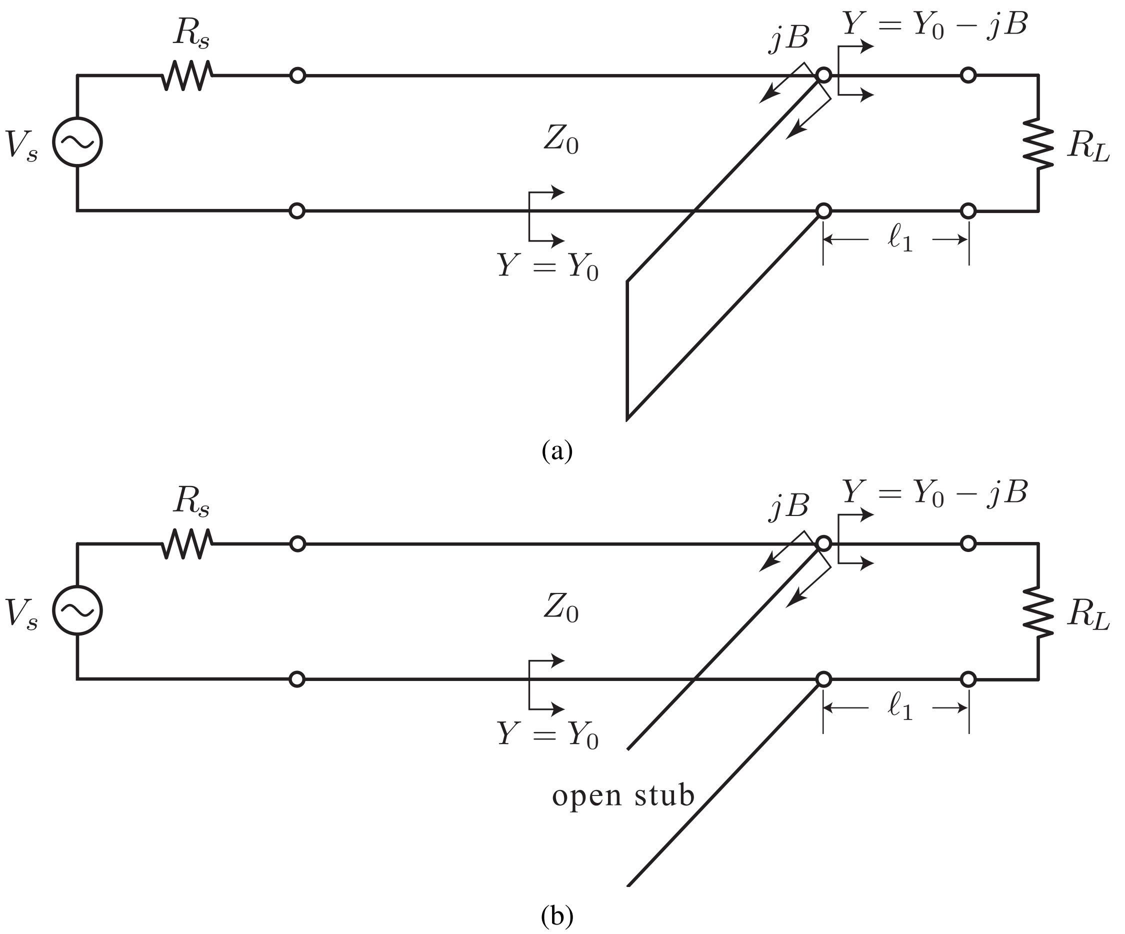


Figure 6.21: A (a) short or open (b) shunt stub can be used in place of the lumped element to obtain the desired imaginary component.

At ℓ_1, the imaginary part of the input admittance is

$$b = \Im(y_1) = \pm \frac{2\rho}{\sqrt{1-\rho^2}} \tag{6.60}$$

Placing a reactance of value $-b$ in shunt provides an impedance match at this particular frequency. If the location of ℓ_1 is not convenient, we can achieve the same result by move back a multiple of $\lambda/2$.

6.3.2 Matching with T-Line Stubs

At high frequencies the matching technique discussed above is difficult due to the lack of lumped passive elements (inductors and capacitors). But short/open pieces of transmission lines simulate fixed reactance over a narrow band. A shorted stub with $\ell < \lambda/4$ looks like an inductor. An open stub with $\ell < \lambda/4$ looks like a capacitor. The procedure is identical to the case with lumped elements but instead of using a capacitor or inductor, we use shorted or open transmission lines. For most transmission line structures, shunt stubs, shown in Fig. 6.21, are easier to fabricate than series stubs. But in theory either shunt or series stubs can be used to obtain the match.

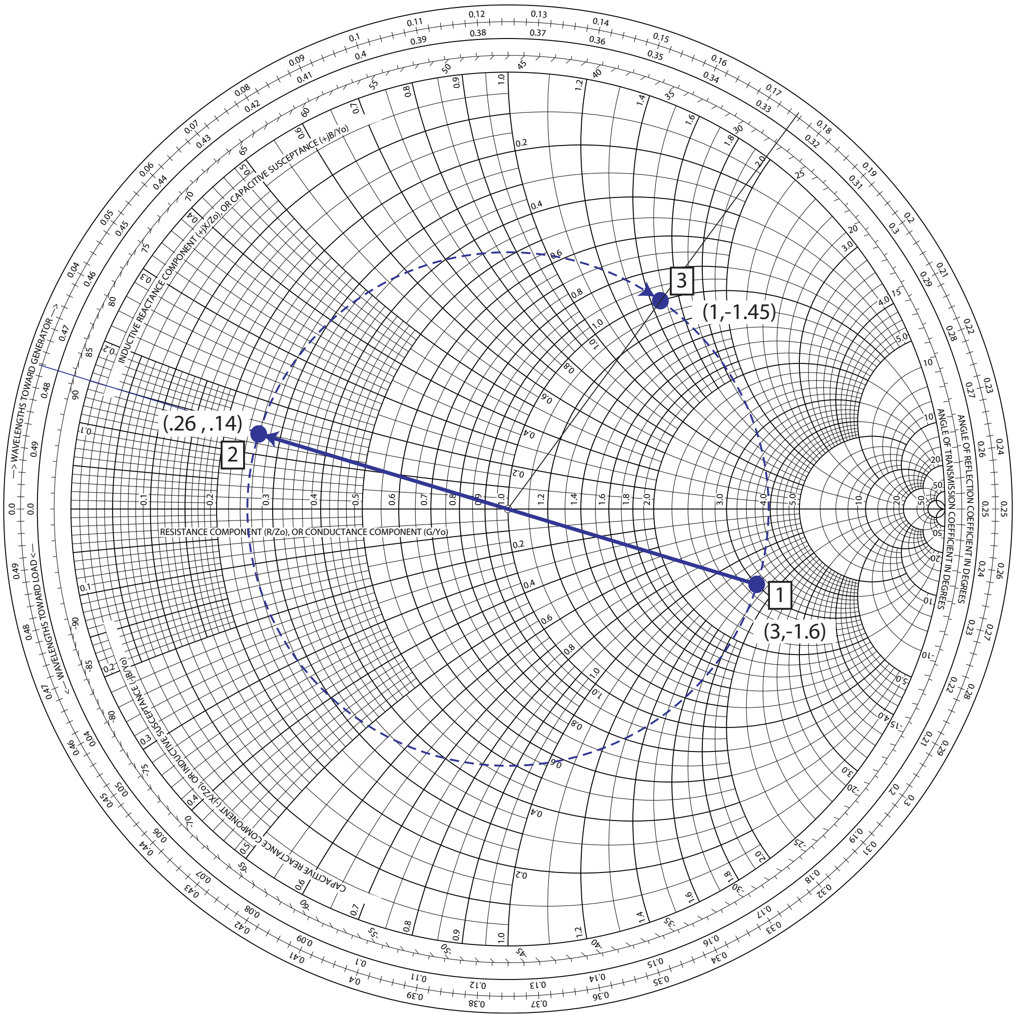


Figure 6.22: Smith Chart calculations for stub matching example.

6.3.3 Matching with the Aid of the Smith Chart

Single stub impedance matching is easy to do with the Smith Chart. Simply find the intersection of the SWR circle with the $r = 1$ circle. The match is at the center of the circle. Grab a reactance in series or shunt to move you there. To do a shunt stub, though, we need to use the admittance chart. To solve the same matching problem with a shunt stub, find the shunt stub value, simply convert the value of $z = 1 + jx$ to $y = 1 + jb$ and place a reactance of $-jb$ in shunt.

Example 7: Consider matching a load impedance $Z_L = 150 - j80$ to Z_0 (usually $50\,\Omega$ at a frequency of 1 GHz.

The normalized impedance $z_L = Z_L/Z_0$ is plotted in Fig. 6.22 and labeled point 1. Since we are connecting a shunt stub, we convert to $y_l = 1/z_L$ by drawing a constant VSWR circle and project through the origin. This point is labeled 2 corresponds to an admittance of $0.26 + j0.14$, which we can read directly from the chart. Now, interpreting the chart as an admittance chart, we move to the $g = 1$ circle along the

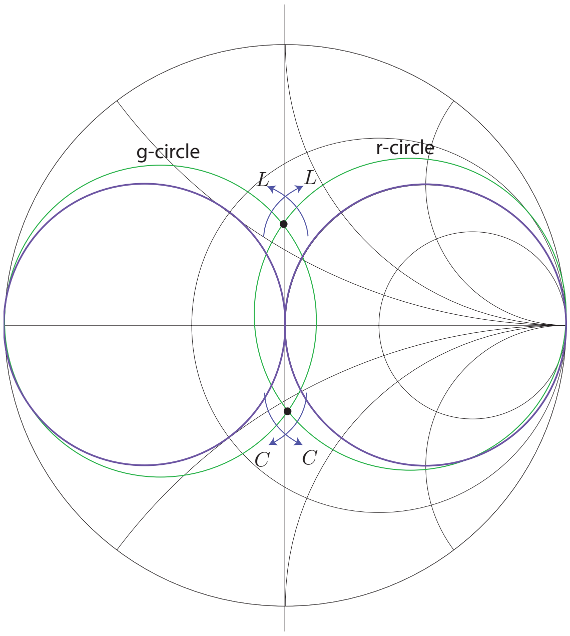


Figure 6.23: Motion along constant r and g circles on the Smith Chart by adding inductors/capacitors.

same constant VSWR circle and arrive at point 3. Reading off the distance traveled on the Smith chart, we have $\ell = \lambda(0.175 - 0.024) = \lambda 0.151$. This distance correspond to an electrical length of about 55°.

At point 2, the normalized admittance is about $1 + 1.45j$. Now, all we need to do is connect a shunt stub with susceptance $-1.45j$ to complete the design. This can be realized as an open or short circuit stub. Since the desired susceptance is negative, the shorter stub is a short circuited "inductive" line of appropriate length. The length of the stub can be calculated directly from

$$y_{stub} = -j \cot \beta \ell$$

or read from the Smith Chart. From the Smith Chart, we start at the short circuit point and travel along the $r = 0$ circle and note that $x = 1/1.45 = .69$ requires a stub length of about 0.096λ, or about 35°. This completes the design of the matching network.

Another great application of the Smith Chart is matching with lumped elements. Previously we described lumped matching networks configurations such as the L, Π, and T matching circuits. Instead of computing element values explicitly, we can use the Smith Chart as a graphical calculator to design matching networks.

6.4 Lumped Matching with the Smith Chart

6.4.1 Lumped Components On Smith Chart

In Section **??** we analyzed "motion" along a transmission line and how it generates circles (an SWR circle) on the Smith Chart. We used this to find a path towards the "matched" circles ($1 + jx$

or $1 + jb$) on the Smith Impedance and Admittance Charts. To finish the matching, we needed to add a series or shunt reactance $-jx$ or $-jb$, taking us to the center of the Smith Chart. In practice, we can use either lumped components or transmission line stubs to achieve this, so it's useful to understand how lumped components move us on the Smith Chart.

It's also important to realize that all of the matching techniques covered in this chapter, including L-match, T-match, and Π-match, can all be done and visualized using the Smith Chart. The advantage of this is that instead of relying on analytical equations, we can explore different matching options step-by-step to realize the best performance (usually lowest Q networks). We also can design matching networks from any complex load / source, without having to do extra "reactance absorbtion" calculations. The end results are the same, but the Smith Chart gives us insights in a visual way.

6.4.2 Lump Component Motion

If we think of an inductor as having any inductance from 0 to ∞, we can imagine that adding an inductor in series to a given load moves us on the constant r circle clock-wise (CW), as shown in Fig. 6.23. To understand why we move CW, consider that adding more inductance increases the reactance, so this should move us towards the higher reactance circles on the right of the chart. Adding a capacitor in series, therefore, should move us CCW, or left.

Similarly, if we add reactance in shunt (parallel), we move along constant g circles. On a combined Z and Y chart, adding inductance in shunt reduces the reactance, so we should move left, or CCW.

To drive these points home, let's do an example. If we start at the bottom half of the Impedance Chart, or start with a capacitive reactive part, then naturally adding an inductor in series cause our path to intersect with the real axis, providing a resonance and a real load. If we continue beyond this point, we then have a net inductive reactance, and occupy the upper half of the unit circle. Our motion is CW because the inductance is making the reactance less negative, moving us towards the left. Once we cross the real axis, the increase in inductance is making us net more positive reactive, moving us now towards the right, in the CW direction. Eventually, as the reactance is increased more and more, we end at the rightmost edge of the unit circle, corresponding to an open circuit due to the infinity reactance.

Since adding a capacitor in series moves counter clock-wise (CCW), we can play out the same scenario if we start at the top of the Smth Chart. Adding more capacitive reactance causes the net reactance to decrease, so we move CCW until we intersect with the the real axis, providing a real load at resonance. Rather than memorize CW and CCW motion, it is more important to reason out the direction of motion in each scenario.

On the Admittance Chart, or the Y-plane, we move on constant g circles. To go clockwise, we add a shunt C, and to move CCW, add a shunt L. For example, if we are in the top half of the Y-plane, we are net capacitive. Adding a shunt C should make us even more capacitive, so the motion should be towards the higher reactance. Adding a shunt L, on the other hands, moves us CCW on a constant g circle towards the real axis, and at the given value we obtain resonance. Further increasing the shunt inductance makes us net negative susceptance, keeping us in the lower half of the plane.

Example 8: Consider again matching a load impedance $Z_L = 150 - j80$ to Z_0 (usually $50\,\Omega$) at a frequency of 1 GHz using lumped elements. From our previous discussion, we know that an L matching network connected in shunt with the output is needed in order to step down from $|Z_L| > Z_0$.

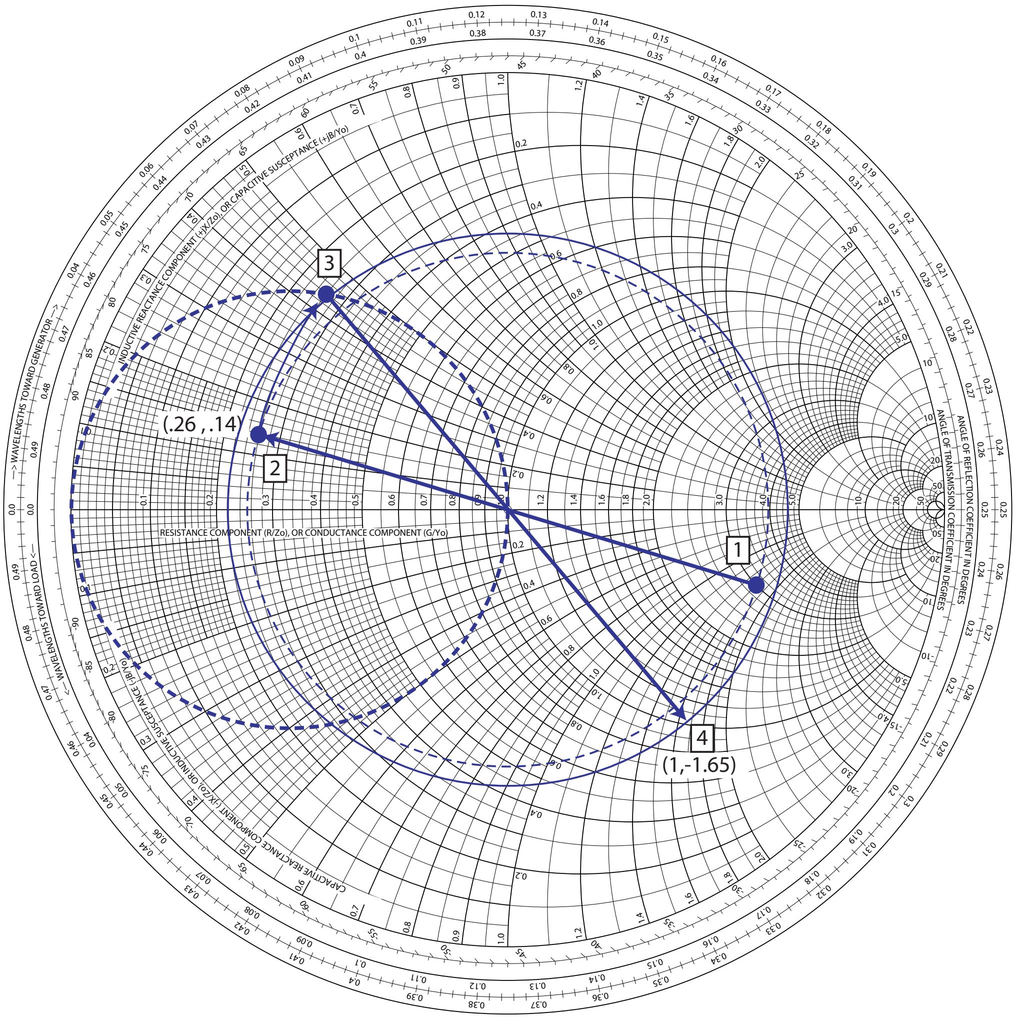


Figure 6.24: Smith Chart calculations for lumped impedance matching network example.

The normalized impedance $z_L = Z_L/Z_0$ is plotted in Fig. 6.24 and labeled point 1. Since we are connecting a reactance in shunt, we convert to $y_l = 1/z_L$ by drawing a constant SWR circle and projecting through the origin. This point is labeled 2 corresponds to an admittance of $0.26 + j0.14$, which we can read directly from the chart. Next we draw the admittance circle $1 + jb$ on the chart, which is a reflection of the $1 + jx$ circle. To move onto this circle, we observe that adding a susceptance of about $0.3j$ moves to a point labeled 3 on the chart. This corresponds to adding a shunt capacitor of value

$$B_C = 0.3/50$$

$$C = \frac{B_C}{\omega} = 955\,\mathrm{fF}$$

Next we convert back to impedance by once again projecting through the origin to arrive at point 4, which corresponds to the normalized impedance of $1 - 1.65j$. It's important to note that this is a consistency check. If we had not ended up on a point $1 + jx$, then there would be an error in our analysis. From here forward it's easy to see that a series reactance of value $+1.65$ will move us to the origin. This corresponds to a series inductor of value

$$X_L = 50 \times 1.65$$

$$L = \frac{X_L}{\omega} = 13\,\mathrm{nH}$$

This completes the design of the matching network. It's interesting to note that we could have traveled to the origin by subtracting reactance from point 2. This would have resulted in another perfectly valid solution to the problem. Generally we would check both solutions and choose the network that best meets our needs (such as bandwidth).

6.4.3 Matching with Lumped Components ($R_L > Z_0$)

We can now generalize and see that an L match can solve any matching problem. The direction of the L depends on our initial point on the Smith Chart. Suppose the load is inside the $1 + jx$ circle, shown in Fig. 6.25. In order to escape from “inside", you cannot move along constant r circles since you'll always stay inside. You must add a shunt reactance, either an inductor as capacitor as shown. On a combined Z/Y chart we can simply follow the trajectory of a constant g circle until we intersect with the $1 + jx$ circle. It's interesting to observe that we can either move up to the $1 + jx$ circle on the top half or move down to the $1 + jx$ bottom half, and both possibilities work. In either case, only a series reactance will be needed to reach the center of the Smith Chart.

If a combined chart is not used, we solve this problem by first converting from z to y. The goal is now to land on the $1 + jb$ constant-g circle. Since we are now in admittance mode, we note that constant g circles intersect with the reflected $1 + jx$ circle at two points. Reflecting these points through the origin using a constant VSWR circle shows the points of intersection with the $1 + jx$ circle in the Z chart. It's confusing, and requires practice, so you may prefer to always use a combined Smith Chart!

In summary, we can say that there are two circuit networks that get to the center, as shown in Fig. 6.26. The difference is one is AC coupled versus DC coupled, so often the application will determine the choice.

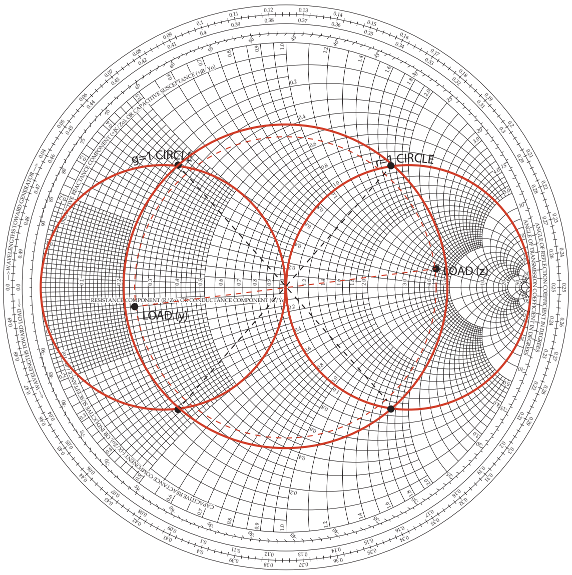


Figure 6.25: When the load is inside the $1+jx$ circle, we must add a shunt reactance to move along a constant g circle in order to reach the $1+jx$ circle. To find the value, we first convert from z to y by reflecting about the origin on the constant SWR circle, then we interpret the chart as a Y chart, reading off the necessary shunt reactance to reach the $1+jb$ circle. It's very confusing, but this is actually the $1+jx$ circle in the admittance chart ! We can see this by reflecting either point of intersection back onto the Z chart $1+jx$ circle.

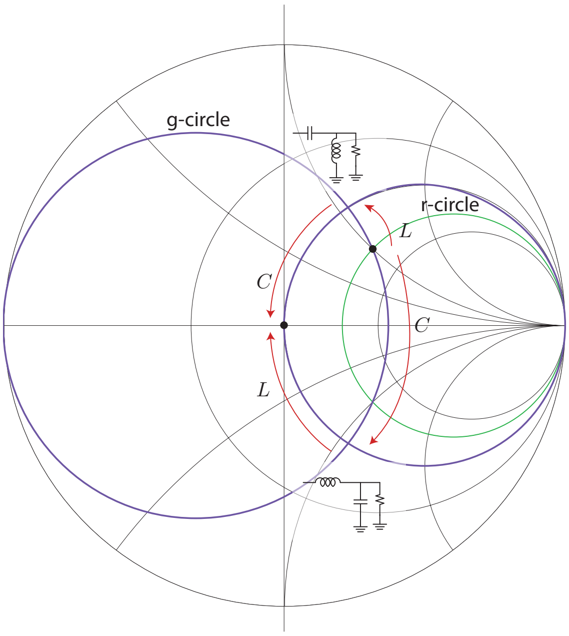


Figure 6.26: Matching networks that move from inside the $1+jx$ circle to the center of the Smith Chart.

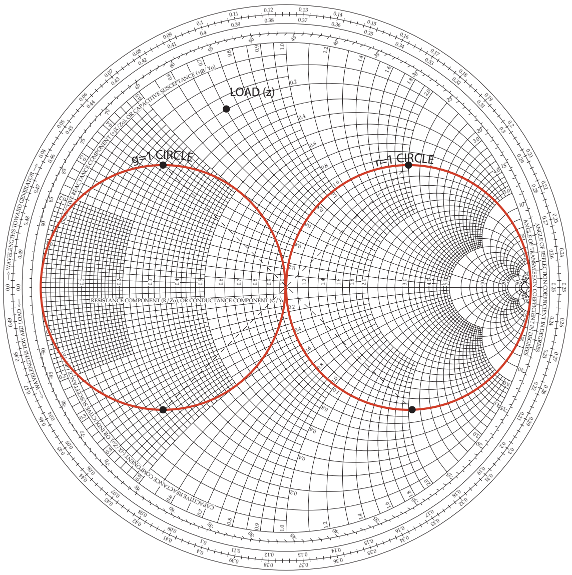


Figure 6.27: Design of a matching network when the load is outside of the $1 + jb$ circle.

Matching with Lumped Components ($R_L < Z_0$)

Suppose the load is outside the $1 + jx$ circle, as shown in Fig. 6.27. This situation is actually much simpler to analyze since we can move along a constant r circle by simply adding a series reactance. The target is the $1 + jb$ circle, since we can then use a shunt element to complete the match.

6.4.4 Q Circles

For designing matching networks, it's very convenient to know the required Q to achieve a match, since the Q is related to bandwidth and loss. Let's find the contours of constant Q on the Smith Chart. Recall that $z = r + jx$ can be related to the reflection coefficient $\rho = u + jv$ by the following equations

$$r = \frac{1 - u^2 - v^2}{(1-u)^2 + v^2} \tag{6.61}$$

$$x = \frac{v(1-u) + v(1+u)}{(1-u)^2 + v^2} \tag{6.62}$$

So that the Q is given by:

$$Q = \frac{x}{r} = \frac{2v}{1 - u^2 - v^2} \tag{6.63}$$

We can write this as

$$Q(1 - u^2 - v^2) = 2v \tag{6.64}$$

$$(1 - u^2) = v^2 + 2v/Q = v^2 + 2v/Q + 1/Q^2 - 1/Q^2 \tag{6.65}$$

$$1 - 1/Q^2 = u^2 + (v + 1/Q)^2 \tag{6.66}$$

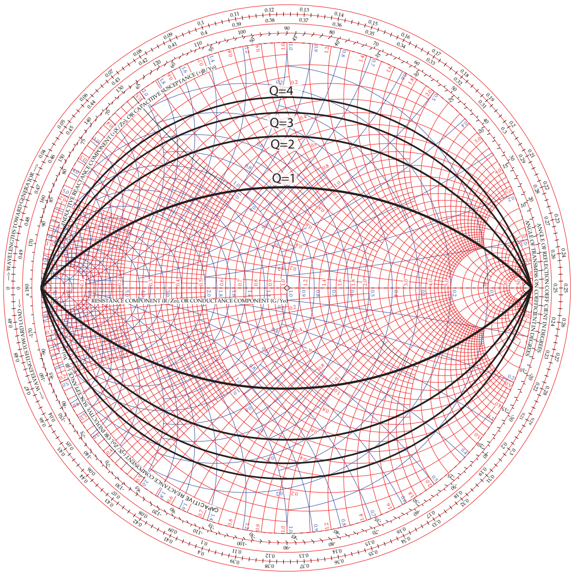


Figure 6.28: Constant Q circles on the Smith Chart. Lower Q is near the real axis.

This is an equation for another circle, centered at $(u,v) = (0,-1/Q)$ with a radius of $R = \sqrt{1+\frac{1}{Q^2}}$. For a capacitive element, the same derivation holds except the circle is centered at $(u,v) = (0,+1/Q)$. Circles of constant Q are plotted in Fig. 6.28. You can hand sketch these circles by noting that $Q = 1$ intersects both the $1+jx$ circle and the $r \pm j$ circle. Likewise, $Q = 2$ will intersect both the $1+jx$ circle and the $r \pm 2j$ circle.

Single Stage vs Two-Stage Matching Network

We can now see visually how a two-stage matching network obtains wider bandwidth and lower loss by reducing the network Q. Notice that with a one stage matching network shown in Fig. 6.29, the Q is fixed (we already knew this but now we see it graphically). On the other hand, with a two-stage matching network we can actually take different paths and lower the Q.

For broadband matching, we should therefore “hug" the real axis $z = r + 0x$. At high frequency, we can also include transmission line leads into the network to make the matching correspond to a more realistic scenario, with distributed parastics. We pick a desired Q (set by bandwidth), and then never allow the matching ratio to go above this value. The number of steps required will be set by the matching ratio.

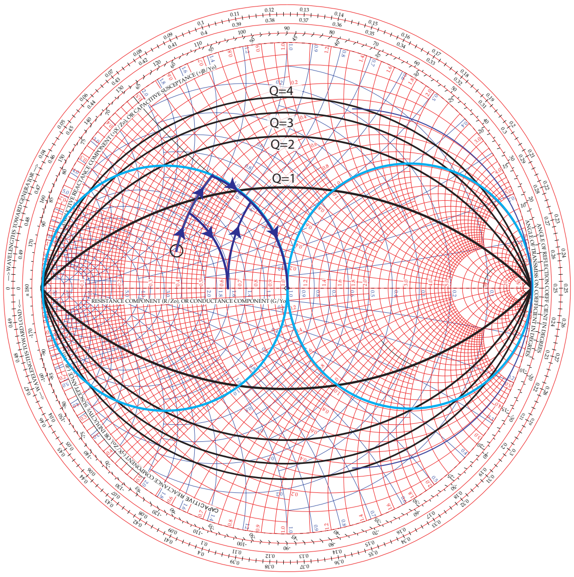
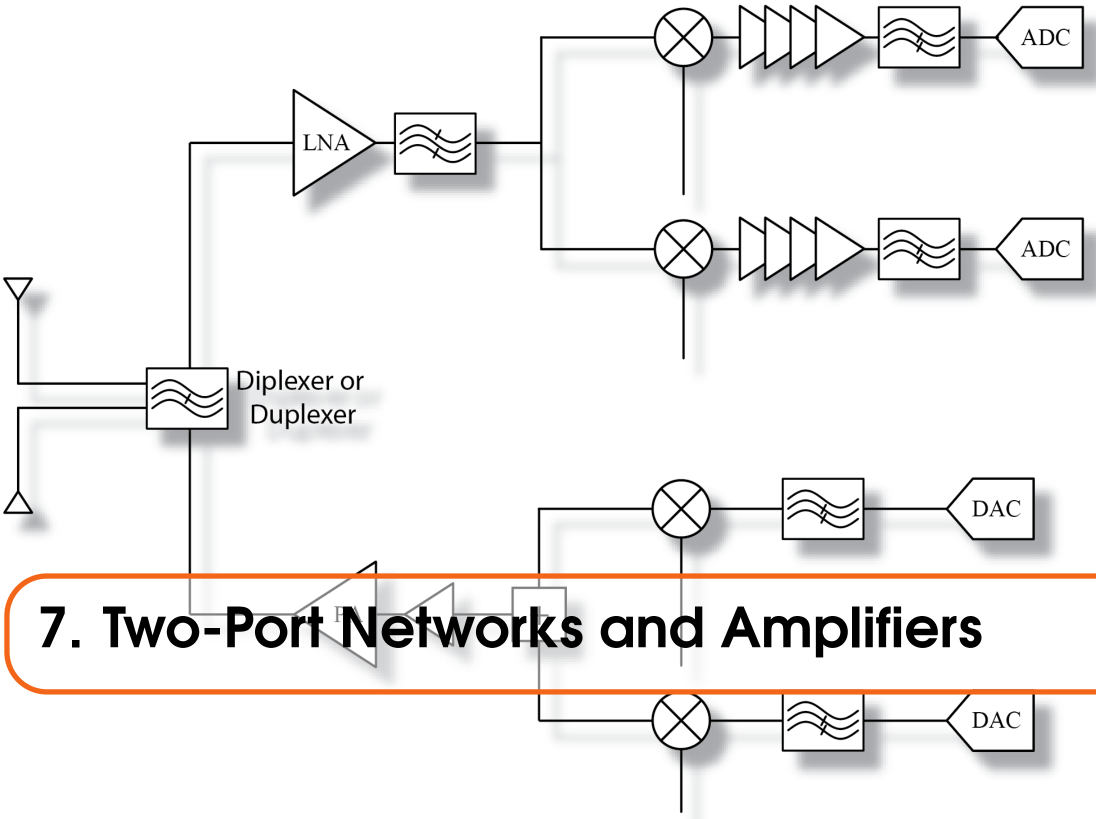


Figure 6.29: Single stage versus multi-stage matching networks.



7. Two-Port Networks and Amplifiers

7.1 Introduction to Two-Port Parameters

Consider the generic two-port amplifier shown in Fig. 7.1. Note that any two-port linear and time-invariant circuit can be described in this way. We can use any two-port parameter set, including admittance parameters Y, impedance parameters Z, hybrid H or inverse-hybrid parameters G. These parameters represent a linear relation between the input/output voltages and currents. If we take linear combinations of current and voltage, we can derive other parameter sets, the most important of which is the scattering or S parameters. We may also choose to represent input versus output, which simplifies analysis of cascade of two-ports, such as the $ABCD$ parameter set

$$\begin{pmatrix} v_1 \\ i_1 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} v_2 \\ -i_2 \end{pmatrix}$$

As shown in Fig. 9.21, the cascade of two blocks is obtained through simple matrix multiplication if we redefine the direction of i_2 so that it flows out of the first block and into the second block.

In this Chapter we review two-port parameters and derive equations for the gain, input/output impedance, and optimal source/load to realize the optimal gain. Next we introduce the important concept of scattering (S) parameters, which are used extensively in high frequency design of amplifiers, filters, and other building blocks. In the laboratory, we measure the properties of a circuit using a network analyzer, which measures the S parameters directly. While it is easy to

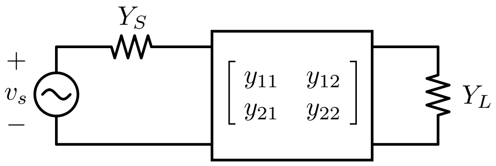


Figure 7.1: A generic amplifier represented as a two-port.

convert from S parameters to other parameters, in many situations it will be convenient to “think” using s-parameters.

7.1.1 Choosing Two-Port Parameters

All two-port parameters are equivalent in their description of a linear system. The best choice of the parameter set is determined by finding the parameters that simplify calculations. For instance, if shunt feedback is applied, Y parameters are most convenient, whereas series feedback favors Z parameters. Other combinations of shunt/series can be easily described by H or G. In Fig. 7.3 the feedback is connected in series with the output and in shunt with the input so we see that we are sensing the output current and feeding back a current to the input. As such the most appropriate parameter set should involve currents/voltages which are the same for both blocks. In this case the input voltage and the output current are the same for each block whereas the total input current and output voltage are a summation of the amplifier and feedback blocks

$$\begin{pmatrix} i_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} i_{a,1} \\ v_{a,2} \end{pmatrix} + \begin{pmatrix} i_{f,1} \\ v_{f,2} \end{pmatrix} = \begin{pmatrix} g^a_{11} & g^a_{12} \\ g^a_{21} & g^a_{22} \end{pmatrix} \begin{pmatrix} v_1 \\ i_2 \end{pmatrix} + \begin{pmatrix} g^f_{11} & g^f_{12} \\ g^f_{21} & g^f_{22} \end{pmatrix} \begin{pmatrix} v_1 \\ i_2 \end{pmatrix} = \begin{pmatrix} g^a_{11}+g^f_{11} & g^a_{12}+g^f_{12} \\ g^a_{21}+g^f_{21} & g^a_{22}+g^f_{22} \end{pmatrix} \begin{pmatrix} v_1 \\ i_2 \end{pmatrix}$$

As mentioned already, the $ABCD$ parameters are useful for cascading two-ports. Many of the results that we derive in terms of say Y-parameters can be applied to other two-port parameters (input impedance, output impedance, gain, etc) by simple substitution. In the laboratory we always use S parameters, since this is actually the way in which we measure two-port parameters at high frequencies.

7.1.2 Y Parameters

First let’s use the Y or admittance parameters since they are familiar and easy to use

$$\begin{pmatrix} i_1 \\ i_2 \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

Notice that y_{11} is the short circuit input admittance

$$y_{11} = \left. \frac{i_1}{v_1} \right|_{v_2=0}$$

The same can be said of y_{22}. The forward transconductance is described by y_{21}

$$y_{21} = \left. \frac{i_2}{v_1} \right|_{v_2=0}$$

whereas the reverse transconductance is described by y_{12}. If a two-port amplifier is unilateral, then $y_{12} = 0$

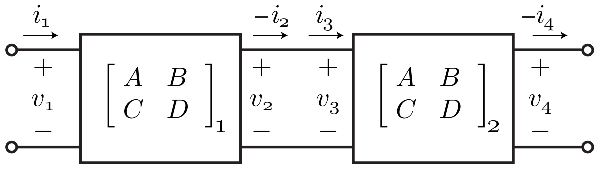

i_1
$-i_2$
i_3
$-i_4$
$+$
v_1
$-$
$\begin{bmatrix} A & B \\ C & D \end{bmatrix}_1$
$+$
v_2
$-$
$+$
v_3
$-$
$\begin{bmatrix} A & B \\ C & D \end{bmatrix}_2$
$+$
v_4
$-$

Figure 7.2: If we reverse the current direction on the second port, we can cascade two-ports using the $ABCD$ parameters.

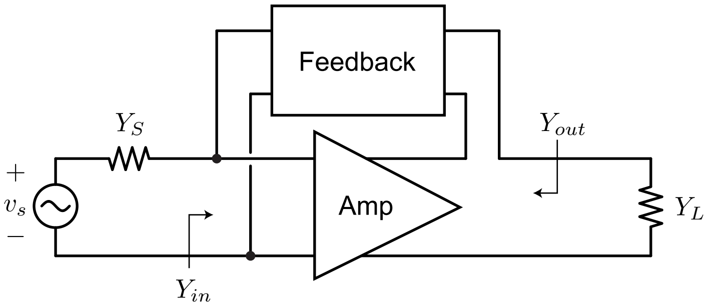


Figure 7.3: A generic feedback amplifier represented as an interconnection of two-ports. Note a series connection is made at the output (current sense) and shunted with the input (current feedback).

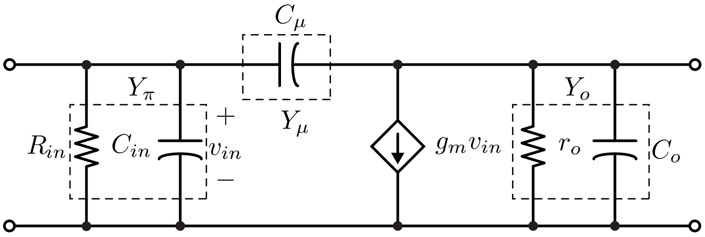


Figure 7.4: A hybrid-pi circuit as a two-port.

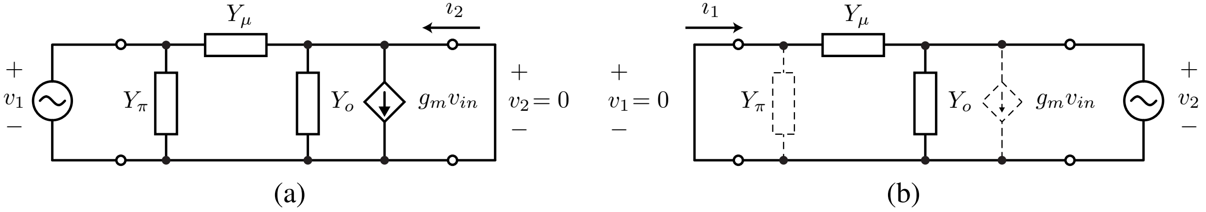


Figure 7.5: Setup to calculate (a) input admittance and (b) output admittance parameters.

7.1.3 Hybrid-Π Admittance Parameters

Let's compute the Y parameters for the common hybrid-Π model shown in Fig. 7.4. With the aid of Fig. 7.5a,

$$y_{11} = y_\pi + y_\mu$$

$$y_{21} = g_m - y_\mu$$

And with the aid of Fig. 7.5b

$$y_{22} = y_o + y_\mu$$

$$y_{12} = -y_\mu$$

Note that the hybrid-Π model is unilateral if $y_\mu = sC_\mu = 0$. Therefore it's unilateral at DC. A good amplifier has a high ratio y_{21}/y_{12} because we expect the forward transconductance to dominate the behavior of the device.

Why Use Two-Port Parameters?

Given that you can analyze amplifiers in detail using KVL/KCL, why use two-port parameters, which are more abstract than the equivalent circuit? The answer is that the parameters are generic and independent of the details of the amplifier. What resides inside the two-port can be a single transistor or a multi-stage amplifier. In addition, high frequency transistors are more easily described by two-port parameters (due to distributed input gate resistance and induced channel resistance). Also, feedback amplifiers can often be decomposed into an equivalent two-port unilateral amplifier and a two-port feedback section. Most importantly, two-port analysis will be used to make some very general conclusions about the stability and "optimal" power gain of a two-port. This in turn will allow us to define some useful metrics for transistors and amplifiers.

7.1.4 Voltage Gain and Input Admittance

Let's begin with some easy calculations for a loaded two-port shown in Fig. 7.1. Since $i_2 = -v_2 Y_L$, we can write

$$(y_{22} + Y_L)v_2 = -y_{21}v_1$$

Which leads to the "internal" two-port gain

$$A_v = \frac{v_2}{v_1} = \frac{-y_{21}}{y_{22} + Y_L}$$

The input admittance is easily calculated from the voltage gain

$$Y_{in} = \frac{i_1}{v_1} = y_{11} + y_{12}\frac{v_2}{v_1}$$

$$Y_{in} = y_{11} - \frac{y_{12}y_{21}}{y_{22} + Y_L}$$

By symmetry we can write down the output admittance by inspection

$$Y_{out} = y_{22} - \frac{y_{12}y_{21}}{y_{11} + Y_S}$$

For a unilateral amplifier $y_{12} = 0$ implies that

$$Y_{in} = y_{11}$$

$$Y_{out} = y_{22}$$

and so the input and output impedance are decoupled. This is a very important property of a unilateral amplifier which simplifies the analysis of optimal gain and stability considerably.

The external voltage gain, or the gain from the voltage source to the output can be derived by a simple voltage divider equation

$$A'_v = \frac{v_2}{v_s} = \frac{v_2}{v_1}\frac{v_1}{v_s} = A_v \frac{Y_S}{Y_{in} + Y_S} = \frac{-Y_S y_{21}}{(y_{22} + Y_L)(Y_S + Y_{in})}$$

If we substitute and simplify the above equation we have

$$A'_v = \frac{-Y_S y_{21}}{(Y_S + y_{11})(Y_L + y_{22}) - y_{12}y_{21}} \tag{7.1}$$

7.1.5 Feedback Amplifiers and Y-Params

Note that in an ideal feedback system, the amplifier is unilateral and the closed loop gain is given by

$$\frac{y}{x} = \frac{A}{1 + Af}$$

If we unilaterize the two-port by arbitrarily setting $y_{12} = 0$, from Eq. 7.1, we have an "open" loop forward gain of

$$A_{vu} = A'_v\big|_{y_{12}=0} = \frac{-Y_S y_{21}}{(Y_S + y_{11})(Y_L + y_{22})}$$

Rewriting the gain A'_v by dividing numerator and denominator by the factor $(Y_S + y_{11})(Y_L + y_{22})$ we have

$$A'_v = \frac{\frac{-Y_S y_{21}}{(Y_S + y_{11})(Y_L + y_{22})}}{1 - \frac{y_{12}y_{21}}{(Y_S + y_{11})(Y_L + y_{22})}}$$

We can now see that the "closed" loop gain with $y_{12} \neq 0$ is given by

$$A'_v = \frac{A_{vu}}{1 + T}$$

where T is identified as the loop gain

$$T = A_{vu} f = \frac{-y_{12}y_{21}}{(Y_S + y_{11})(Y_L + y_{22})}$$

Using the last equation also allows us to identify the feedback factor

$$f = \frac{y_{12}}{Y_S}$$

If we include the loading by the source Y_S, the input admittance of the amplifier is given by

$$Y_{in} = Y_S + y_{11} - \frac{y_{12}y_{21}}{Y_L + y_{22}}$$

Note that this can be re-written as

$$Y_{in} = (Y_S + y_{11})\left(1 - \frac{y_{12}y_{21}}{(Y_S + y_{11})(Y_L + y_{22})}\right)$$

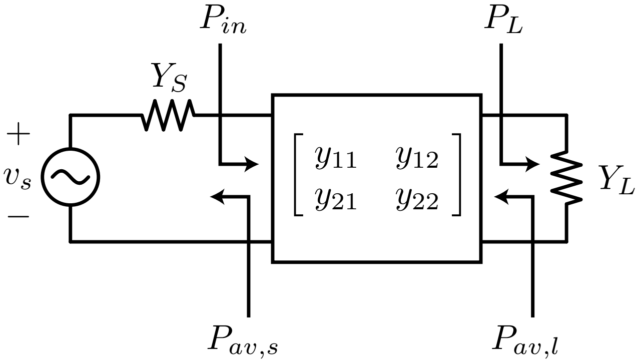


Figure 7.6: Various definitions of power in a two-port.

The last equation can be re-written as

$$Y_{in} = (Y_S + y_{11})(1+T)$$

Since $Y_S + y_{11}$ is the input admittance of a unilateral amplifier, we can interpret the action of the feedback as raising the input admittance by a factor of $1+T$. Likewise, the same analysis yields

$$Y_{out} = (Y_L + y_{22})(1+T)$$

It's interesting to note that the same equations are valid for series feedback using Z parameters, in which case the action of the feedback is to boost the input and output impedance. For the hybrid H parameters, the action of the series feedback at the input also raises the input impedance but the action of the shunt output connection lowers the output impedance. The inverse applies for the inverse-hybrid G parameters.

7.2 Power Gain

We can define power gain in many different ways. You may think that the *power gain* G_p is defined as follows

$$G_p = \frac{P_L}{P_{in}} = f(Y_L, Y_{ij}) \neq f(Y_S)$$

is the best way, but notice that this gain is a function of the load admittance Y_L and the two-port parameters Y_{ij}, but not the source admittance. In other words, G_p is the load power normalized by the input power. If the input power is very small, such as in a source mismatch condition, then the output power will also be small. This is hidden from G_p.

The *transducer gain* defined by

$$G_T = \frac{P_L}{P_{av,S}} = f(Y_L, Y_S, Y_{ij})$$

measures the power deliverd to the load normalized by the available power from the source ($P_{av,S}$). This is a measure of the efficacy of the two-port as it compares the power at the load to a simple conjugate match. As such it is a function of the source and the load.

The *available power gain* is defined as follows

$$G_a = \frac{P_{av,L}}{P_{av,S}} = f(Y_S, Y_{ij}) \neq f(Y_L)$$

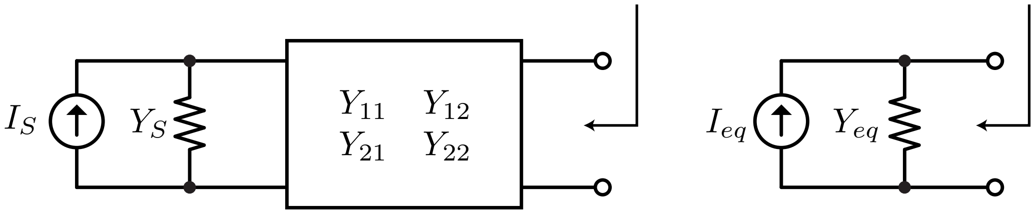


Figure 7.7: The Norton equivalent of a two-port from the output port.

where the available power from the two-port is denoted $P_{av,L}$. This quantity is only a function of the load admittance and measures the efficiency of the output matching network.

The power gain is readily calculated from the input admittance and voltage gain

$$P_{in} = \frac{|V_1|^2}{2}\Re(Y_{in})$$

$$P_L = \frac{|V_2|^2}{2}\Re(Y_L)$$

$$G_p = \left|\frac{V_2}{V_1}\right|^2 \frac{\Re(Y_L)}{\Re(Y_{in})}$$

$$G_p = \frac{|Y_{21}|^2}{|Y_L+Y_{22}|^2}\frac{\Re(Y_L)}{\Re(Y_{in})}$$

To derive the available power gain, consider a Norton equivalent for the two-port where (short port two) shown in Fig. 7.7

$$I_{eq} = I_2 = Y_{21}V_1 = \frac{Y_{21}}{Y_{11}+Y_S}I_S$$

The Norton equivalent admittance is simply the output admittance of the two-port

$$Y_{eq} = Y_{22} - \frac{Y_{21}Y_{12}}{Y_{11}+Y_S}$$

The available power at the source and load are given by

$$P_{av,S} = \frac{|I_S|^2}{8\Re(Y_S)}$$

$$P_{av,L} = \frac{|I_{eq}|^2}{8\Re(Y_{eq})}$$

$$G_a = \left|\frac{I_{eq}}{I_S}\right|^2 \frac{\Re(Y_S)}{\Re(Y_{eq})}$$

$$G_a = \left|\frac{Y_{21}}{Y_{11}+Y_S}\right|^2 \frac{\Re(Y_S)}{\Re(Y_{eq})}$$

The transducer gain is given by

$$G_T = \frac{P_L}{P_{av,S}} = \frac{\frac{1}{2}\Re(Y_L)|V_2|^2}{\frac{|I_S|^2}{8\Re(Y_S)}} = 4\Re(Y_L)\Re(Y_S)\left|\frac{V_2}{I_S}\right|^2$$

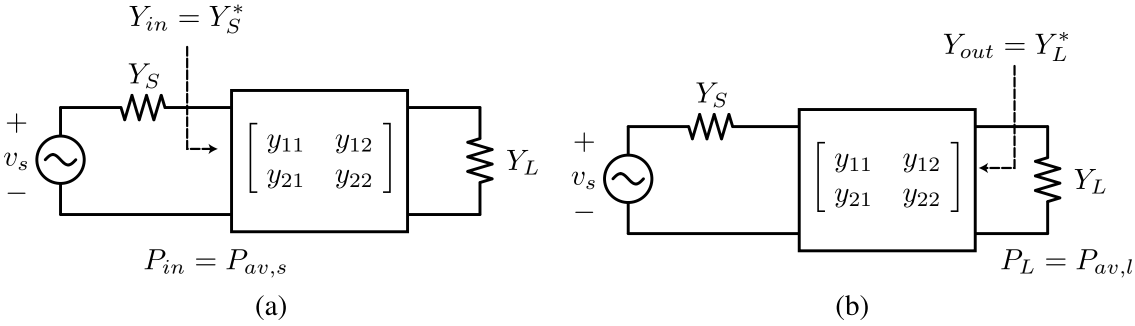


Figure 7.8: (a) A two-port matched at the input port. (b) A two-port matched at the output port.

We need to find the output voltage in terms of the source current. Using the voltage gain we have and input admittance we have

$$\left|\frac{V_2}{V_1}\right| = \left|\frac{Y_{21}}{Y_L+Y_{22}}\right|$$

$$I_S = V_1(Y_S+Y_{in})$$

$$\left|\frac{V_2}{I_S}\right| = \left|\frac{Y_{21}}{Y_L+Y_{22}}\right|\frac{1}{|Y_S+Y_{in}|}$$

$$|Y_S+Y_{in}| = \left|Y_S+Y_{11}-\frac{Y_{12}Y_{21}}{Y_L+Y_{22}}\right|$$

We can now express the output voltage as a function of source current as

$$\left|\frac{V_2}{I_S}\right|^2 = \frac{|Y_{21}|^2}{|(Y_S+Y_{11})(Y_L+Y_{22})-Y_{12}Y_{21}|^2}$$

And thus the transducer gain

$$G_T = \frac{4\Re(Y_L)\Re(Y_S)|Y_{21}|^2}{|(Y_S+Y_{11})(Y_L+Y_{22})-Y_{12}Y_{21}|^2}$$

There is no need to redefine the power gains for the other parameter sets since *all* of the gain expression we have derived are in the exact same form for the impedance, hybrid, and inverse hybrid matrices. Simply change y to z, h or g.

7.2.1 Comparison of Power Gains

Since $P_{in} \le P_{av,s}$, we see that $G_T \le G_p$. Under what condition is $G_T = G_p$? Simply when the input impedance is conjugately matches to the source impedance (Fig. 7.8a). Since $P_L \le P_{av,l}$, we see that $G_T \le G_a$. Again, equality is obtained when the load is conjugately matched to the two-port output impedance (Fig. 7.8b). In summary

$$G_{T,max,L} = \frac{P_L(Y_L = Y_{out}^*)}{P_{av,S}} = G_a$$

$$G_{T,max,S} = G_T(Y_{in} = Y_S^*) = G_p$$

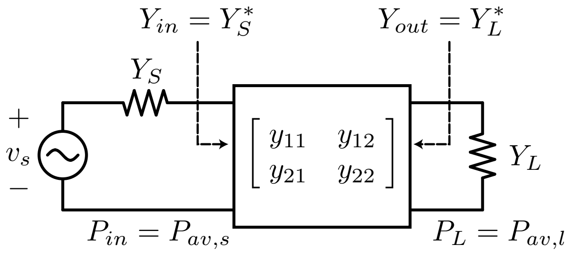


Figure 7.9: The bi-conjugate match, or simultaneous input and output match.

Input and Output Conjugate Match

It should be clear now that if we simultaneously conjugate match *both* the input and output of a two-port, we'll obtain the maximum possible power gain (Fig. 7.9). Under this condition all three gains are equal

$$G_{p,max} = G_{T,max} = G_{a,max}$$

This is thus the recipe for calculating the optimal source and load impedance in to maximize gain

$$Y_{in} = Y_{11} - \frac{Y_{12}Y_{21}}{Y_L + Y_{22}} = Y_S^*$$

$$Y_{out} = Y_{22} - \frac{Y_{12}Y_{21}}{Y_S + Y_{11}} = Y_L^*$$

Solution of the above four equations (real/imag) results in the optimal $Y_{S,opt}$ and $Y_{L,opt}$, or the solution to a pair of quadratic equations.

Calculation of Optimal Source/Load

Another approach to the problem of calculating the optimal source/load is to simply equate the partial derivatives of G_T with respect to the source/load admittance to zero to

$$\frac{\partial G_T}{\partial G_S} = \frac{\partial G_T}{\partial B_S} = \frac{\partial G_T}{\partial G_L} = \frac{\partial G_T}{\partial B_L} = 0$$

Again we have four equations. But we should be smarter about this and recall that the maximum gains are all equal. Since G_a and G_p are only a function of the source or load, we can get away with only solving two equations. For instance

$$\frac{\partial G_a}{\partial G_S} = \frac{\partial G_a}{\partial B_S} = 0$$

This yields $Y_{S,opt}$ and by setting $Y_L = Y_{out}^*$ we can find the $Y_{L,opt}$. Likewise we can also solve

$$\frac{\partial G_p}{\partial G_L} = \frac{\partial G_p}{\partial B_L} = 0$$

And now use $Y_{S,opt} = Y_{in}^*$. Let's outline the procedure for the optimal power gain. We'll use the power gain G_p and take partials with respect to the load. Let

$$Y_{jk} = m_{jk} + jn_{jk}$$

$$Y_L = G_L + jX_L$$

$$Y_{12}Y_{21} = P + jQ = Le^{j\phi}$$

$$G_p = \frac{|Y_{21}|^2}{D} G_L$$

$$\Re\left(Y_{11} - \frac{Y_{12}Y_{21}}{Y_L + Y_{22}}\right) = m_{11} - \frac{\Re(Y_{12}Y_{21}(Y_L + Y_{22})^*)}{|Y_L + Y_{22}|^2}$$

$$D = m_{11}|Y_L + Y_{22}|^2 - P(G_L + m_{22}) - Q(B_L + n_{22})$$

$$\frac{\partial G_p}{\partial B_L} = 0 = -\frac{|Y_{21}|^2 G_L}{D^2} \frac{\partial D}{\partial B_L}$$

Solving the above equation we arrive at the following solution

$$B_{L,opt} = \frac{Q}{2m_{11}} - n_{22}$$

In a similar fashion, solving for the optimal load conductance

$$G_{L,opt} = \frac{1}{2m_{11}} \sqrt{(2m_{11}m_{22} - P)^2 - L^2}$$

If we substitute these values into the equation for G_p (lot's of algebra ...), we arrive at

$$G_{p,max} = \frac{|Y_{21}|^2}{2m_{11}m_{22} - P + \sqrt{(2m_{11}m_{22} - P)^2 - L^2}}$$

Notice that for the solution to exists, G_L must be a real number. In other words

$$(2m_{11}m_{22} - P)^2 > L^2$$

$$(2m_{11}m_{22} - P) > L$$

$$K = \frac{2m_{11}m_{22} - P}{L} > 1$$

The condition on the factor K is important as we will later show that it also corresponds to an unconditionally stable two-port. We can recast all of the work up to here in terms of K

$$Y_{S,opt} = \frac{Y_{12}Y_{21} + |Y_{12}Y_{21}|(K + \sqrt{K^2 - 1})}{2\Re(Y_{22})}$$

$$Y_{L,opt} = \frac{Y_{12}Y_{21} + |Y_{12}Y_{21}|(K + \sqrt{K^2 - 1})}{2\Re(Y_{11})}$$

$$G_{p,max} = G_{T,max} = G_{a,max} = \frac{Y_{21}}{Y_{12}} \frac{1}{K + \sqrt{K^2 - 1}}$$

7.2.2 Maximum Gain

The maximum gain is usually written in the following insightful form

$$G_{max} = \frac{Y_{21}}{Y_{12}}(K - \sqrt{K^2 - 1})$$

For a reciprocal network, such as a passive element, $Y_{12} = Y_{21}$ and thus the maximum gain is given by the second factor

$$G_{r,max} = K - \sqrt{K^2 - 1}$$

Since $K > 1$, $|G_{r,max}| < 1$. The reciprocal gain factor is known as the efficiency of the reciprocal network. The first factor, on the other hand, is a measure of the non-reciprocity.

The case of a unilateral amplifier is of particular interest

$$G_{TU} = \frac{4|y_{21}|^2\Re(Y_L)\Re(Y_S)}{|(y_{22}+Y_L)(Y_S+Y_{in})|^2}$$

The transducer gain is maximum under a conjugate input/output match

$$Y_S = Y_{in}^* = Y_{11}^*$$

$$Y_L = Y_{out}^* = Y_{22}^*$$

Resulting in a maximum unilateral gain

$$G_{TU,max} = \frac{|y_{21}|^2}{4\Re(Y_L)\Re(Y_S)}$$

Take for instance the hybrid-π model discussed earlier (Fig. 7.24). If we assume the model is unilateral, e.g. $C_\mu \approx 0$, then

$$y_{11} = Y_\pi + Y_\mu \approx Y_\pi$$

$$y_{22} = Y_o + Y_\mu \approx Y_o$$

$$y_{21} = gm - Y_\mu \approx g_m$$

$$y_{11} = Y_\mu \approx 0$$

Using the formula derived for $G_{TU,max}$ we have

$$G_{TU,max} = \frac{4g_m^2}{\Re(y_{11})\Re(y_{22})}$$

For an ideal FET, the input admittance is imaginary, e.g. $\Re(y_{11}) = 0$, which implies infinite power gain. This is a non-physical result and so we can see that a real FET must have physical resistance on the input side. In practice the gate resistance comes from the poly-gate structure, the interconnect, and the induced channel resistance.

7.2.3 Two-Port Stability and Negative Resistance

A two-port network is unstable if it supports non-zero currents/voltages with passive terminations

$$\begin{pmatrix} i_1 \\ i_2 \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

Since $i_1 = -v_1 Y_S$ and $i_2 = -v_2 Y_L$

$$\begin{pmatrix} y_{11}+Y_S & y_{12} \\ y_{21} & y_{22}+Y_L \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0$$

The only way to have a non-trial solution is for the determinant of the matrix to be zero at a particular frequency. Taking the determinant of the matrix we have

$$(Y_S + y_{11})(Y_L + y_{22}) - y_{12}y_{21} = 0$$

Let's re-write the above in the following form

$$Y_S + y_{11} - \frac{y_{12}y_{21}}{y_{22}+Y_L} = 0$$

or

$$Y_S + Y_{in} = 0$$

equivalently

$$Y_L + Y_{out} = 0$$

A network is unstable at a particular frequency if $Y_S + Y_{in} = 0$, which means the condition is satisfied for both the real and imaginary part. In particular

$$\Re(Y_S + Y_{in}) = \Re(Y_S) + \Re(Y_{in}) = 0$$

Since the terminations are passive, $\Re(Y_S) > 0$ which implies that

$$\Re(Y_{in}) < 0$$

The same equations also show that

$$\Re(Y_{out}) < 0$$

So if these conditions are satisfied, the two-port is unstable.

The conditions for stability are a function of the source and load termination

$$\Re(Y_{in}) = \Re\left(y_{11} - \frac{y_{12}y_{21}}{Y_L + y_{22}}\right) > 0$$

$$\Re(Y_{out}) = \Re\left(y_{22} - \frac{y_{12}y_{21}}{Y_S + y_{11}}\right) > 0$$

For a unilateral amplifier, the conditions are simple and only depend on the two-port

$$\Re(y_{11}) > 0$$

$$\Re(y_{22}) > 0$$

Stability Factor

In general, it can be shown that a two-port is absolutely stable if

$$\Re(y_{11}) > 0$$

$$\Re(y_{22}) > 0$$

and

$$K > 1$$

Where the stability factor K is given by

$$K = \frac{2\Re(y_{11})\Re(y_{22}) - \Re(y_{12}y_{21})}{|y_{12}y_{21}|}$$

The stability of a unilateral amplifier with $y_{12} = 0$ is infinite ($K = \infty$) which implies absolute stability (as long as $\Re(y_{11}) > 0$ and $\Re(y_{22}) > 0$). An amplifier with absolute stability means that the two-port is stable for all passive terminations at either the load or the source. This is a conservative situation in applications where the source and load impedances are well specified and well controlled. But in certain situations the load or source impedance may vary greatly. For instance the input impedance of an antenna can vary if the antenna is moved in proximity to conductors, bent, shorted, or broken. An unstable two-port can be stabilized by adding sufficient loss at the input or output to overcome the negative conductance.

7.3 Scattering Parameters

Voltages and currents are difficult to measure directly at microwave frequencies. The Z matrix requires "opens", and it's hard to create an ideal open circuit due to parasitic capacitance and radiation. Likewise, a Y matrix requires "shorts", again ideal short circuits are impossible at high frequency due to the finite inductance. Furthermore, many active devices could oscillate under the open or short termination. In practice, we measure scattering or S-parameters at high frequency. The measurement is direct and only involves measurement of relative quantities (such as the standing wave ratio). It's important to realize that although we associate S parameters with high frequency and wave propagation, the concept is valid for any frequency.

7.3.1 Power Flow in an One-Port

The concept of scattering parameters is very closely related to the concept of power flow. For this reason, we begin with the simple observation that the power flow into a one-port circuit can be written in the following form

$$P_{in} = P_{avs} - P_r$$

where P_{avs} is the available power from the source. Unless otherwise stated, let us assume sinusoidal steady-staet. If the source has a real resistance of Z_0, this is simply given by

$$P_{avs} = \frac{V_s^2}{8Z_0}$$

Of course if the one-port is conjugately matched to the source, then it will draw the maximal available power from the source. Otherwise, the power P_{in} is always less than P_{avs}, which is reflected in our equation. In general, P_r represents the wasted or untapped power that one-port circuit is "reflecting" back to the source due to a mismatch. For passive circuits it's clear that each term in the equation is positive and $P_{in} \geq 0$.

The complex power absorbed by the one-port is given by

$$P_{in} = \frac{1}{2}(v_1 \cdot I_1^* + V_1^* \cdot i_1)$$

which allows us to write

$$P_r = P_{avs} - P_{in} = \frac{V_s^2}{4Z_0} - \frac{1}{2}(V_1 I_1^* + V_1^* I_1)$$

the factor of 4 instead of 8 is used since we are now dealing with complex power. The average power can be obtained by taking one half of the real component of the complex power. If the one-port has an input impedance of Z_{in}, then the power P_{in} is expanded to

$$P_{in} = \frac{1}{2}\left(\frac{Z_{in}}{Z_{in}+Z_0}V_s \cdot \frac{V_s^*}{(Z_{in}+Z_0)^*} + \frac{Z_{in}^*}{(Z_{in}+Z_0)^*}V_s \cdot \frac{V_s}{(Z_{in}+Z_0)}\right)$$

which is easily simplified to

$$P_{in} = \frac{|V_s|^2}{2Z_0}\left(\frac{Z_0 Z_{in} + Z_{in}^* Z_0}{|Z_{in}+Z_0|^2}\right)$$

where we have assumed Z_0 is real. With the exception of a factor of 2, the premultiplier is simply the source available power, which means that our overall expression for the reflected power is given by

$$P_r = \frac{V_s^2}{4Z_0}\left(1 - 2\frac{Z_0 Z_{in} + Z_{in}^* Z_0}{|Z_{in}+Z_0|^2}\right)$$

which can be simplified

$$P_r = P_{avs}\left|\frac{Z_{in}-Z_0}{Z_{in}+Z_0}\right|^2 = P_{avs}|\Gamma|^2$$

where we have defined Γ, or the reflection coefficient, as

$$\Gamma = \frac{Z_{in}-Z_0}{Z_{in}+Z_0}$$

From the definition it is clear that $|\Gamma| \leq 1$, which is just a re-statement of the conservation of energy implied by our assumption of a passive load. This constant Γ, also called the scattering parameter of a one-port, plays a very important role. On one hand we see that it is has a one-to-one relationship with Z_{in}. Given Γ we can solve for Z_{in} by inverting the above equation

$$Z_{in} = Z_0\frac{1+\Gamma}{1-\Gamma}$$

which means that all of the information in Z_{in} is also in Γ. Moreover, since $|\Gamma| < 1$, we see that the space of the semi-infinite space of all impedance values with real positive components (the right-half plane) maps into the unit circle. This is a great compression of information which allows us to visualize the entire space of realizable impedance values by simply observing the unit circle. We shall find wide application for this concept when finding the appropriate load/source impedance for an amplifier to meet a given noise or gain specification.

More importantly, Γ expresses very direct and obviously the power flow in the circuit. If $\Gamma = 0$, then the one-port is absorbing all the possible power available from the source. If $|\Gamma| = 1$ then the one-port is not absorbing any power, but rather "reflecting" the power back to the source. Clearly an open circuit, short circuit, or a reactive load cannot absorb net power. For an open and short load, this is obvious from the definition of Γ. For a reactive load, this is pretty clear if we substitute $Z_{in} = jX$

$$|\Gamma_X| = \left|\frac{jX-Z_0}{jX+Z_0}\right| = \left|\frac{\sqrt{X^2+Z_0^2}}{\sqrt{X^2+Z_0^2}}\right| = 1$$

The transformation between impedance and Γ is a well known mathematical transform (see Bilinear Transform). It is a conformal mapping (meaning that it preserves angles) which maps vertical and horizontal lines in the impedance plane into circles. We have already seen that the jX axis is mapped onto the unit circle.

Since $|\Gamma|^2$ represents power flow, we may imagine that Γ should represent the flow of voltage, current, or some linear combination thereof. Consider taking the square root of the basic equation we have derived

$$\sqrt{P_r} = \Gamma\sqrt{P_{avs}}$$

where we have retained the positive root. We may write the above equation as

$$b_1 = \Gamma a_1$$

where a and b have the units of square root of power and represent signal flow in the network. How are a and b related to currents and voltage? Let

$$a_1 = \frac{V_1+Z_0I_1}{2\sqrt{Z_0}}$$

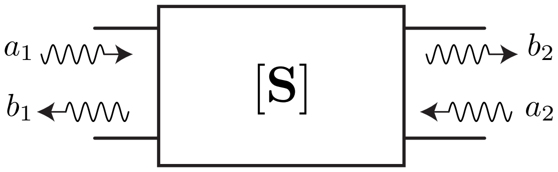


Figure 7.10: A two-port black box with normalized waves a and b.

and

$$b_1 = \frac{V_1 - Z_0 I_1}{2\sqrt{Z_0}}$$

It is now easy to show that for the one-port circuit, these relations indeed represent the available and reflected power:

$$|a_1|^2 = \frac{|V_1|^2}{4Z_0} + \frac{Z_0|I_1|^2}{4} + \frac{V_1^* \cdot I_1 + V_1 \cdot I_1^*}{4}$$

Now substitute $V_1 = Z_{in}V_s/(Z_{in}+Z_0)$ and $I_1 = V_s/(Z_{in}+Z_0)$ we have

$$|a_1|^2 = \frac{|V_s|^2}{4Z_0}\frac{|Z_{in}|^2}{|Z_{in}+Z_0|^2} + \frac{Z_0|V_s|^2}{4|Z_{in}+Z_0|^2} + \frac{|V_s|^2}{4Z_0}\frac{Z_{in}^*Z_0 + Z_{in}Z_0}{|Z_{in}+Z_0|^2}$$

or

$$|a_1|^2 = \frac{|V_s|^2}{4Z_0}\left(\frac{|Z_{in}|^2 + Z_0^2 + Z_{in}^*Z_0 + Z_{in}Z_0}{|Z_{in}+Z_0|^2}\right) = \frac{|V_s|^2}{4Z_0}\left(\frac{|Z_{in}+Z_0|^2}{|Z_{in}+Z_0|^2}\right) = P_{avs}$$

In a like manner, the square of b is given by many similar terms

$$|b_1|^2 = \frac{|V_s|^2}{4Z_0}\left(\frac{|Z_{in}|^2 + Z_0^2 - Z_{in}^*Z_0 - Z_{in}Z_0}{|Z_{in}+Z_0|^2}\right) = P_{avs}\left|\frac{Z_{in}-Z_0}{Z_{in}+Z_0}\right|^2 = P_{avs}|\Gamma|^2$$

as expected. We can now see that the expression $b = \Gamma \cdot a$ is analogous to the expression $V = Z \cdot I$ or $I = Y \cdot V$ and so it can be generalized to an N-port circuit. In fact, since a and b are linear combinations of v and i, there is a one-to-one relationship between the two. Taking the sum and difference of a and b we arrive at

$$a_1 + b_1 = \frac{2V_1}{2\sqrt{Z_0}} = \frac{V_1}{\sqrt{Z_0}}$$

which is related to the port voltage and

$$a_1 - b_1 = \frac{2Z_0 I_1}{2\sqrt{Z_0}} = \sqrt{Z_0} I_1$$

which is related to the port current.

7.3.2 Scattering Parameters for a Two-Port

Let us now generalize the concept of scattering parameters to a two-port and write

$$b_1 = S_{11}a_1 + S_{12}a_2$$

$$b_2 = S_{21}a_1 + S_{22}a_2$$

with reference to Fig. 7.10, we can interpret the above equation as follows. If we drive a two-port with a source, then a_1 represents the available power from the source, and some fraction of that power will be reflected due to S_{11} (mismatch at the input) and some fraction of that power will "transmitted" to the the second port. In other words, the signal b_2 represents the transmitted signal flowing into the load connected on port two. But if port two is not matched, then this power cannot be fully absorbed and some of that power must flow back into the system, represented by a_2. Let us make this intuitive picture more rigorous by finding the meaning of each parameter. First consider S_{11}, which is easy to to understand if we can set $a_2 = 0$. From the definition of a_2, we have

$$a_2 = \frac{V_2 + Z_0 I_2}{2\sqrt{Z_0}} = 0$$

or

$$\frac{V_2}{-I_2} = Z_0$$

which is tantamount to loading the second port with a resistance of Z_0. Under this condition, then, we can readily identify S_{11}

$$S_{11} = \left.\frac{b_1}{a_1}\right|_{a_2=0}$$

as simply the same as Γ for a one-port circuit. In other words, this is the ratio of the signal "reflected" back to the source and $1 - |S_{11}|^2$ therefore represents the amount of the available source power flowing into the two-port circuit. Note that this is true as long as the second port is terminated in Z_0. Using the second equation, we have

$$S_{21} = \left.\frac{b_2}{a_1}\right|_{a_2=0}$$

which represents the signal flowing *out* of the two-port and towards the load normalized by the available source power flowing into port 1. In other words, this represents the gain of the two-port under the matched condition. Note that under matched conditions the signals a_1 and b_2 take on particular simply forms

$$a_1 = \frac{V_1 + I_1 Z_0}{2\sqrt{Z_0}} = V\frac{1 + \frac{I_1}{V_1} Z_0}{2\sqrt{Z_0}} = \frac{2V_1}{\sqrt{Z_0}}$$

and

$$b_2 = \frac{V_2 - I_2 Z_0}{2\sqrt{Z_0}} = V_2\frac{1 - \frac{I_1}{V_1} Z_0}{2\sqrt{Z_0}} = \frac{2V_2}{\sqrt{Z_0}}$$

which means

$$S_{21} = \frac{V_2}{V_1} = 2\frac{V_2}{V_s}$$

which is simply twice the voltage gain of the circuit from the load to the source. This follows since the signal V_1 is exactly half of the source voltage under matched conditions. $|S_{21}|^2$ is the power gain of the two-port when both ports are terminated by Z_0 since in this case all the available source power flows into the two port and the amount appearing at the load is given by $|b_2|^2$. If $|S_{21}| > 1$, that means there is more power at the load than power flowing into the two-port, which can only be true if the two-port is active. If we interchange the order of the ports, we immediately see that S_{22} is likewise the output reflection coefficient under matched conditions and S_{12} is the reverse gain of the two-port.

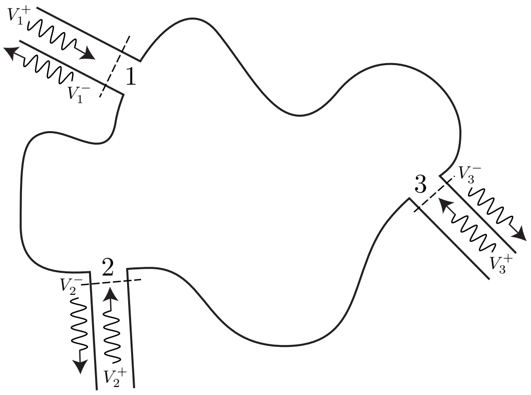


Figure 7.11: An arbitrary N port circuit with incident and reflected waves at each port.

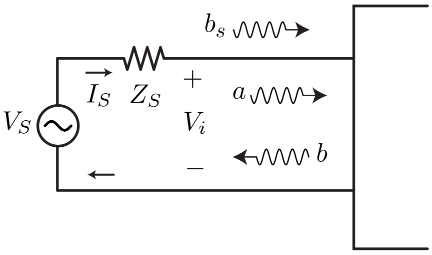


Figure 7.12: A voltage source with source impedance Z_S.

7.3.3 Representation of Source

How do we represent the voltage source in Fig. 7.12 with a source impedance $Z_s \neq Z_0$ directly with S parameters? Start with the I-V relation

$$V_i = V_s - I_s Z_s$$

The voltage source can be represented directly for s-parameter analysis as follows. First note that

$$(a+b)\sqrt{Z_0} = V_s + \left(\frac{a-b}{\sqrt{Z_0}}\right) Z_s$$

or

$$b(Z_0 - Z_s) = \sqrt{Z_0}V_s - a(Z_s + Z_0)$$

Solve these equations for a, the power flowing into a two-port

$$a = \frac{\sqrt{Z_0}V_s}{Z_s + Z_0} + b\frac{Z_0 - Z_s}{Z_0 + Z_s}$$

Define Γ_s as the source reflection coefficient and b_s as the source signal

$$\Gamma_s = \frac{Z_0 - Z_s}{Z_0 + Z_s}$$

$$b_s = \frac{\sqrt{Z_0}V_s}{Z_s + Z_0}$$

With these definitions in place, the power flow away from the source has a simple form

$$a = b_s + b\Gamma_s$$

If the source is matched to Z_0, then $\Gamma_s = 0$ and the total power flowing out of the source is the same as the source power. Otherwise the source signal power should include any reflections occurring at the source itself.

Available Power from Source

A useful quantity is the available power from a source under conjugate matched conditions. Let's begin by noting that the power flowing into a load Γ_L is given by

$$P_L = |a|^2 - |b|^2 = |a|^2(1 - |\Gamma_L|^2)$$

Using the fact that $b = \Gamma_L a$, the input power signal is given by

$$a = b_s + b\Gamma_s = b_s + \Gamma_L\Gamma_s a$$

or

$$a = \frac{b_s}{1 - \Gamma_L\Gamma_s}$$

Therefore the power flowing into the load is given by

$$P_L = \frac{|b_s|^2(1 - |\Gamma_L|^2)}{|1 - \Gamma_L\Gamma_s|^2}$$

To draw the available power from the source, we should conjugately match the load $\Gamma_L = \Gamma_s^*$

$$P_{avs} = P_L|_{\Gamma_L = \Gamma_s^*} = \frac{|b_s|^2(1 - |\Gamma_s|^2)}{|1 - |\Gamma_s|^2|^2} = \frac{|b_s|^2}{1 - |\Gamma_s|^2}$$

7.3.4 Incident and Scattering Waves

If you're familiar with transmission line theory, then you clearly understand the origin of the term "reflected" signal and "transmitted" signal. In transmission line theory, signal a is often called the "forward" wave and represetned by v^+ and b is called the reflected or scattered wave and denoted by v^-. In a transmission line the power is actually reflected since the source does not know the port impedance until information travels from the source to the two-port and then back to the source again (limited by the speed of light) and so there is a physical origin to the terminology. In lumped circuit theory, there is no time delay, but we use the same terminology. For an N port circuit, consider N transmission line connected to each port (Fig. 7.11) and define the reference plane as the point where the transmission line terminates onto the port. In transmission line parlance, these signals are voltages (and currents), so we define them as follows

$$v^+ = V + IZ_0$$

$$v^- = V - IZ_0$$

Notice the similarity to the definition of a and b, where the normalization and power factors are missing. The vectors v^- and v^+ are the incident and "scattered" waveforms

$$v^+ = \begin{pmatrix} V_1^+ \\ V_2^+ \\ V_3^+ \\ \vdots \end{pmatrix}$$

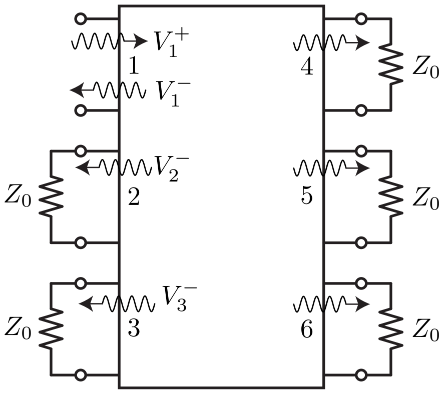


Figure 7.13: An N port circuit with all ports terminated so that $V_j^+ = 0$ for $j \neq 1$.

$$v^- = \begin{pmatrix} V_1^- \\ V_2^- \\ V_3^- \\ \vdots \end{pmatrix}$$

Because the N port is linear, we expect that scattered field to be a linear function of the incident field

$$v^- = Sv^+$$

S is the scattering matrix

$$S = \begin{pmatrix} S_{11} & S_{12} & \cdots \\ S_{21} & \ddots & \\ \vdots & & \end{pmatrix}$$

The fact that the S matrix exists can be easily proved using transmission line theory. The voltage and current on each transmission line termination can be written as

$$V_i = V_i^+ + V_i^-$$

$$I_i = Y_0(I_i^+ - I_i^-)$$

Inverting these equations

$$V_i + Z_0 I_i = V_i^+ + V_i^- + V_i^+ - V_i^- = 2V_i^+$$

$$V_i - Z_0 I_i = V_i^+ + V_i^- - V_i^+ + V_i^- = 2V_i^-$$

Thus v^+,v^- are simply linear combinations of the port voltages and currents. By the uniqueness theorem, then, $v^- = Sv^+$.

Measurement of S_{ij}

The term S_{ij} can be computed directly by the following formula

$$S_{ij} = \left. \frac{V_i^-}{V_j^+} \right|_{V_k^+ = 0 \forall k \neq j}$$

Solve for $V_k^+ = 0$

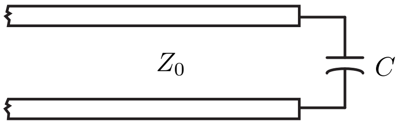
$$V_k^+ = V_k + I_k Z_0 = 0$$

or

$$\frac{V_k}{-I_k} = Z_0$$

which means we terminate port k with an impedance Z_0 and measure the scattered waves. From a transmission line perspective, to measure S_{ij}, drive port j with a wave amplitude of V_j^+ and terminate all other ports with the characteristic impedance of the lines (so that $V_k^+ = 0$ for $k \neq j$), as shown in Fig. 7.13. Then observe the wave amplitude coming out of the port i.

Example 9:



Let's calculate the S parameter for a capacitor

$$S_{11} = \frac{V_1^-}{V_1^+}$$

We can also do the calculation directly from the definition of S parameters. Substituting for the current in a capacitor

$$V_1^- = V - IZ_0 = V - j\omega CV = V(1 - j\omega CZ_0)$$

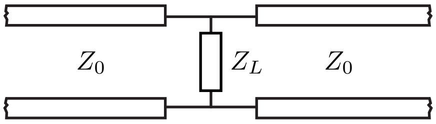
$$V_1^+ = V + IZ_0 = V + j\omega CV = V(1 + j\omega CZ_0)$$

Alternatively, this is just the reflection coefficient for a capacitor

$$S_{11} = \rho_L = \frac{Z_C - Z_0}{Z_C + Z_0} = \frac{\frac{1}{j\omega C} - Z_0}{\frac{1}{j\omega C} + Z_0}$$

$$= \frac{1 - j\omega CZ_0}{1 + j\omega CZ_0}$$

and the ratio yields the same result as expected.



Example 10:

Consider a shunt impedance connected at the junction of two transmission lines. If we terminate port 2 in an impedance Z_0, then the current $I_1 = V_1/R||Z_0$, which allows us to write

$$V_1^- = V_1 - I_1 Z_0 = V_1\left(1 - \frac{Z_0}{R||Z_0}\right)$$

In a like manner, the incident wave is given by

$$V_1^+ = V_1 + I_1 Z_0 = V_1\left(1 + \frac{Z_0}{R||Z_0}\right)$$

The ratio gives us the scattering coefficient

$$S_{11} = \frac{1 - \frac{Z_0}{R||Z_0}}{1 + \frac{Z_0}{R||Z_0}} = \frac{R||Z_0 - Z_0}{R||Z_0 + Z_0}$$

From transmission line theory, we recognize this to be the reflection coefficient seen at port one when port two is terminated in Z_0. We can also calculate S_{21} by noting that

$$V_2^- = V_2 - Z_0 I_2 = V_1 - Z_0\left(\frac{-V_1}{Z_0}\right) = 2V_1$$

Taking the ratio with the incident wave V_1^+

$$S_{21} = \left.\frac{V_2^-}{V_1^+}\right|_{V_2^- = 0} = \frac{2}{1 + \frac{Z_0}{R||Z_0}} = \frac{2R||Z_0}{R||Z_0 + Z_0}$$

By symmetry, we have the complete two-port scattering parameters. Another approach is to use transmission line theory. Start by observing that the voltage at the junction is continuous. The currents, though, differ

$$V_1 = V_2$$

$$I_1 + I_2 = Y_L V_2$$

To compute S_{11}, enforce $V_2^+ = 0$ by terminating the line. Thus we can be re-write the above equations

$$V_1^+ + V_1^- = V_2^-$$

$$Y_0(V_1^+ - V_1^-) = Y_0 V_2^- + Y_L V_2^- = (Y_L + Y_0)V_2^-$$

We can now solve the above equation for the reflected and transmitted wave

$$V_1^- = V_2^- - V_1^+ = \frac{Y_0}{Y_L + Y_0}(V_1^+ - V_1^-) - V_1^+$$

$$V_1^-(Y_L+Y_0+Y_0) = (Y_0-(Y_0+Y_L))V_1^+$$

$$S_{11} = \frac{V_1^-}{V_1^+} = \frac{Y_0-(Y_0+Y_L)}{Y_0+(Y_L+Y_0)} = \frac{Z_0||Z_L - Z_0}{Z_0||Z_L+Z_0}$$

The above equation can be written by inspection since $Z_0||Z_L$ is the effective load seen at the junction of port 1. Thus for port 2 we can write

$$S_{22} = \frac{Z_0||Z_L - Z_0}{Z_0||Z_L+Z_0}$$

Likewise, we can solve for the transmitted wave, or the wave scattered into port 2

$$S_{21} = \frac{V_2^-}{V_1^+}$$

Since $V_2^- = V_1^+ + V_1^-$, we have

$$S_{21} = 1 + S_{11} = \frac{2Z_0||Z_L}{Z_0||Z_L+Z_0}$$

By symmetry, we can deduce S_{12} as

$$S_{12} = \frac{2Z_0||Z_L}{Z_0||Z_L+Z_0}$$

Conversion Formula

Since V^+ and V^- are related to V and I, it's easy to find a formula to convert for Z or Y to S

$$V_i = V_i^+ + V_i^- \rightarrow v = v^+ + v^-$$

$$Z_{i0}I_i = V_i^+ - V_i^- \rightarrow Z_0 i = v^+ - v^-$$

Now starting with $v = Zi$, we have

$$v^+ + v^- = ZZ_0^{-1}(v^+ - v^-)$$

Note that Z_0 is the scalar port impedance

$$v^-(I + ZZ_0^{-1}) = (ZZ_0^{-1} - I)v^+$$

$$v^- = (I + ZZ_0^{-1})^{-1}(ZZ_0^{-1} - I)v^+ = Sv^+$$

We now have a formula relating the Z matrix to the S matrix

$$S = (ZZ_0^{-1} + I)^{-1}(ZZ_0^{-1} - I) = (Z + Z_0 I)^{-1}(Z - Z_0 I)$$

Recall that the reflection coefficient for a load is given by the same equation!

$$\overline{\rho} = \frac{Z/Z_0 - 1}{Z/Z_0 + 1}$$

To solve for Z in terms of S, simply invert the relation

$$Z_0^{-1}ZS + IS = Z_0^{-1}Z - I$$

$$Z_0^{-1}Z(I - S) = S + I$$

$$Z = Z_0(I + S)(I - S)^{-1}$$

As expected, these equations degenerate into the correct form for a 1 × 1 system

$$Z_{11} = Z_0 \frac{1 + S_{11}}{1 - S_{11}}$$

Reciprocal Networks

We have found that the Z and Y matrix are symmetric. Now let's see what we can infer about the S matrix.

$$v^+ = \frac{1}{2}(v + Z_0 i)$$

$$v^- = \frac{1}{2}(v - Z_0 i)$$

Substitute $v = Zi$ in the above equations

$$v^+ = \frac{1}{2}(Zi + Z_0 i) = \frac{1}{2}(Z + Z_0)i$$

$$v^- = \frac{1}{2}(Zi - Z_0 i) = \frac{1}{2}(Z - Z_0)i$$

Since $i = i$, the above equation must result in consistent values of i

$$2(Z + Z_0)^{-1} v^+ = 2(Z - Z_0)^{-1} v^-$$

Thus

$$S = (Z - Z_0)(Z + Z_0)^{-1}$$

Consider the transpose of the S matrix

$$S^t = \left((Z + Z_0)^{-1}\right)^t (Z - Z_0)^t$$

Recall that Z_0 is a diagonal matrix

$$S^t = (Z^t + Z_0)^{-1}(Z^t - Z_0)$$

If $Z^t = Z$ (reciprocal network), then we have

$$S^t = (Z + Z_0)^{-1}(Z - Z_0)$$

Previously we found that

$$S = (Z + Z_0)^{-1}(Z - Z_0)$$

So that we see that the S matrix is also symmetric (under reciprocity)

$$S^t = S$$

To see this another way, note that in effect we have shown that

$$(Z + I)^{-1}(Z - I) = (Z - I)(Z + I)^{-1}$$

This is easy to demonstrate if we note that

$$Z^2 - I = Z^2 - I^2 = (Z + I)(Z - I) = (Z - I)(Z + I)$$

In general matrix multiplication does not commute, but here it does

$$(Z - I) = (Z + I)(Z - I)(Z + I)^{-1}$$

$$(Z + I)^{-1}(Z - I) = (Z - I)(Z + I)^{-1}$$

Thus we see that $S^t = S$.

Scattering Parameters of a Lossless Network

Consider the total power dissipated by a lossless network (must sum to zero)

$$P_{av} = \frac{1}{2}\Re\left(v^t i^*\right) = 0$$

Expanding in terms of the wave amplitudes

$$= \frac{1}{2}\Re\left((v^+ + v^-)^t Z_0^{-1}(v^+ - v^-)^*\right)$$

where we assume that Z_0 are real numbers and equal. The notation is about to get ugly in the expansion

$$= \frac{1}{2Z_0}\Re\left(v^{+^t}v^{+^*} - v^{+^t}v^{-^*} + v^{-^t}v^{+^*} - v^{-^t}v^{-^*}\right)$$

The middle terms sum to a purely imaginary number. Let $x = v^+$ and $y = v^-$

$$y^t x^* - x^t y^* = y_1 x_1^* + y_2 x_2^* + \cdots - x_1 y_1^* + x_2 y_2^* + \cdots = a - a^*$$

We have shown that

$$P_{av} = \frac{1}{2Z_0}\left(\underbrace{v^{+^t}v^{+}}_{\text{total incident power}} - \underbrace{v^{-^t}v^{-^*}}_{\text{total reflected power}}\right) = 0$$

This is a rather obvious result. It simply says that the incident power is equal to the reflected power (because the N port is lossless). Since $v^- = Sv^+$

$$v^{+^t}v^{+} = (Sv^+)^t(Sv^+)^* = v^{+^t}S^tS^*v^{+^*}$$

This can only be true if S is a unitary matrix

$$S^tS^* = I$$

$$S^* = (S^t)^{-1}$$

Orthogonal Properties of S

Expanding out the matrix product

$$\delta_{ij} = \sum_k (S^T)_{ik}S^*_{kj} = \sum_k S_{ki}S^*_{kj}$$

For $i = j$ we have

$$\sum_k S_{ki}S^*_{ki} = 1$$

For $i \neq j$ we have

$$\sum_k S_{ki}S^*_{kj} = 0$$

The dot product of any column of S with the conjugate of that column is unity while the dot product of any column with the conjugate of a different column is zero. If the network is reciprocal, then $S^t = S$ and the same applies to the rows of S. Note also that $|S_{ij}| \leq 1$.

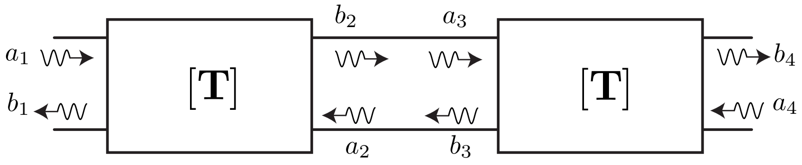


Figure 7.14: The cascade of two two-ports. The incident and reflected power at the connection point.

Shift in Reference Planes

A convenient feature of the scattering parameters is that we can easily move the reference plane. In other words, if we connect transmission lines of arbitrary length to any port, we can easily de-embed their effect. We'll derive a new matrix S' related to S. Let's call the waves at the new reference ν

$$v^- = Sv^+$$

$$\nu^- = S'\nu^+$$

Since the waves on the lossless transmission lines only experience a phase shift, we have a phase shift of $\theta_i = \beta_i \ell_i$

$$\nu_i^- = v^- e^{-j\theta_i}$$

$$\nu_i^+ = v^+ e^{j\theta_i}$$

Or we have

$$\begin{bmatrix} e^{j\theta_1} & 0 & \cdots & \\ 0 & e^{j\theta_2} & \cdots & \\ 0 & 0 & e^{j\theta_3} & \cdots \\ \vdots & & & \end{bmatrix} \nu^- = S \begin{bmatrix} e^{-j\theta_1} & 0 & \cdots & \\ 0 & e^{-j\theta_2} & \cdots & \\ 0 & 0 & e^{-j\theta_3} & \cdots \\ \vdots & & & \end{bmatrix} \nu^+$$

So we see that the new S matrix is simply

$$S' = \begin{bmatrix} e^{-j\theta_1} & 0 & \cdots & \\ 0 & e^{-j\theta_2} & \cdots & \\ 0 & 0 & e^{-j\theta_3} & \cdots \\ \vdots & & & \end{bmatrix} S \begin{bmatrix} e^{-j\theta_1} & 0 & \cdots & \\ 0 & e^{-j\theta_2} & \cdots & \\ 0 & 0 & e^{-j\theta_3} & \cdots \\ \vdots & & & \end{bmatrix}$$

7.3.5 Scattering Transfer Parameters

Up to now we found it convenient to represent the scattered waves in terms of the incident waves. But what if we wish to cascade two ports as shown in Fig. 7.14? Since b_2 flows into a_1', and likewise b_1' flows into a_2, would it not be convenient if we defined the a relationship between a_1,b_1 and b_2,a_2? In other words we have

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} b_2 \\ a_2 \end{bmatrix}$$

Notice carefully the order of waves (a,b) in reference to the figure above. This allows us to cascade matrices

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix} = T_1 \begin{bmatrix} b_2 \\ a_2 \end{bmatrix} = T_1 \begin{bmatrix} a_3 \\ b_3 \end{bmatrix} = T_1 T_2 \begin{bmatrix} b_4 \\ a_4 \end{bmatrix}$$

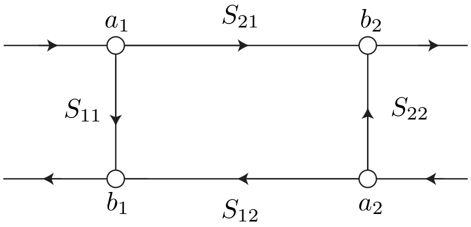


Figure 7.15: The signal-flow graph of a two-port.

7.4 Signal-Flow Analysis

Signal-flow analysis is a technique for graphically calculating the transfer function directly using scattering parameters. Each signal a and b in the system is represented by a node. Branches connect nodes with “strength” given by the scattering parameter. For example, a general two-port is represented in Fig. 7.15. Using three simple rules, we can simplify signal flow graphs to the point that detailed calculations are done by inspection. Of course we can always “do the math” using algebra, so pick the technique that you like best.

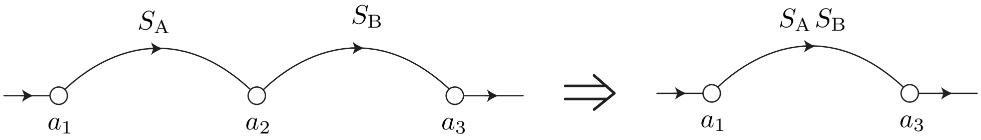


Figure 7.16: The series connection rule.

- Rule 1: (series rule) By inspection of Fig. 7.16, we have the cascade.

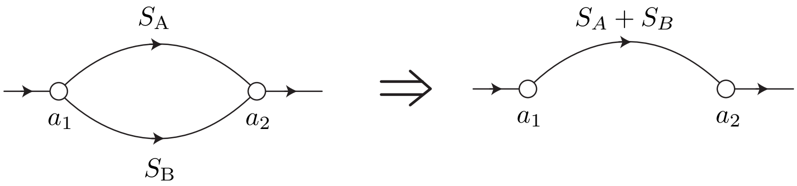


Figure 7.17: The parallel connection rule.

- Rule 2: (parallel rule) Clear by inspection of Fig. 7.17.
- Rule 3: (self-loop rule) We can remove a “self-loop” in Fig. 7.18 by multiplying branches feeding the node by $1/(1-S_B)$ since

$$a_2 = S_A a_1 + S_B a_2$$

$$a_2(1-S_B) = S_A a_1$$

$$a_2 = \frac{S_A}{1-S_B} a_1$$

- Rule 4: (splitting rule) We can duplicate node a_2 in Fig. 7.19 by splitting the signals at an earlier phase

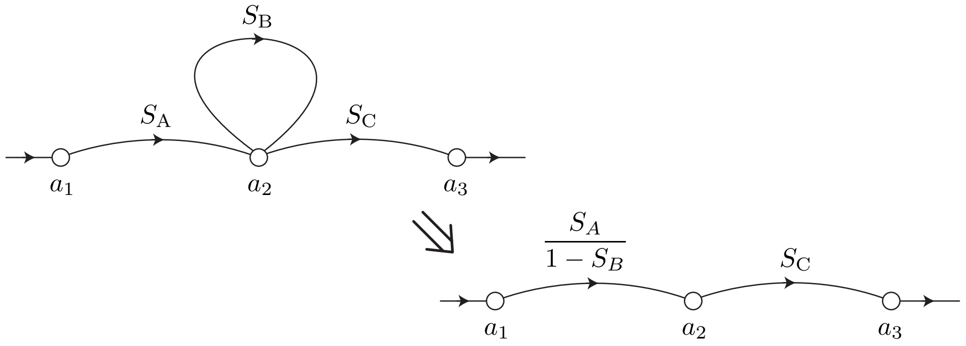


Figure 7.18: The self-loop elimination rule.

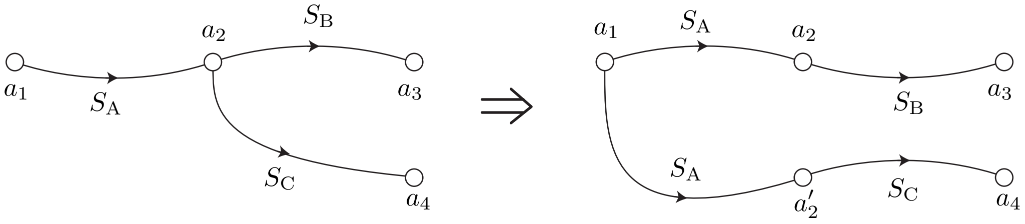


Figure 7.19: The splitting rule.

Using the above rules, we can calculate the input reflection coefficient of a two-port terminated by $\Gamma_L = b_1/a_1$ shown in Fig. 7.20a using a couple of steps. First we notice that there is a self-loop around b_2 (Fig. 7.20b). Next we remove the self loop and from here it's clear that the (Fig. 7.20c)

$$\Gamma_{in} = \frac{b_1}{a_1} = S_{11} + \frac{S_{21}S_{12}\Gamma_L}{1-S_{22}\Gamma_L}$$

7.4.1 Mason's Rule

Using Mason's Rule, you can calculate the transfer function for a signal flow graph by "inspection"

$$T = \frac{P_1\left(1-\sum\mathscr{L}(1)^{(1)}+\sum\mathscr{L}(2)^{(1)}-\cdots\right)+P_2\left(1-\sum\mathscr{L}(1)^{(2)}+\cdots\right)+\cdots}{1-\sum\mathscr{L}(1)+\sum\mathscr{L}(2)-\sum\mathscr{L}(3)+\cdots}$$

Each P_i defines a *path*, a directed route from the input to the output not containing each node more than once. The value of P_i is the product of the branch coefficients along the path. For instance, in Fig. 7.21a, the path from b_s to b_1 ($T = b_1/b_s$) has two paths, $P_1 = S_{11}$ and $P_2 = S_{21}\Gamma_L S_{12}$

Loop of Order Summation Notation

The notation $\sum\mathscr{L}(1)$ is the sum over all first order loops. A "first order loop" is defined as product of the branch values in a loop in the graph. For the example shown in Fig. 7.21b, we have $\Gamma_s S_{11}$, $S_{22}\Gamma_L$, and $\Gamma_s S_{21}\Gamma_L S_{12}$. A "second order loop" $\mathscr{L}(2)$ is the product of two non-touching first-order loops. For instance, since loops $S_{11}\Gamma_s$ and $S_{22}\Gamma_L$ do not touch, their product is a second order loop. A "third order loop" $\mathscr{L}(3)$ is likewise the product of three non-touching first order loops. The notation $\sum\mathscr{L}(1)^{(p)}$ is the sum of all first-order loops that do not touch the path p. For path P_1, we have $\sum\mathscr{L}(1)^{(1)} = \Gamma_L S_{22}$ but for path P_2 we have $\sum\mathscr{L}(1)^{(2)} = 0$.

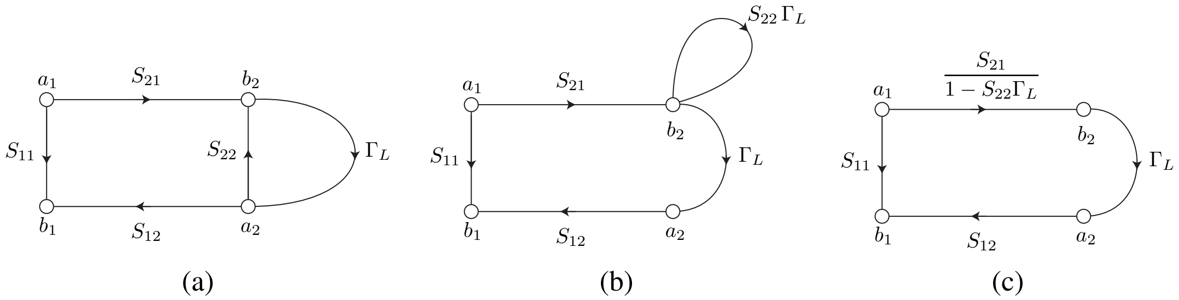


Figure 7.20: (a) A two-port terminated in a load Γ_L. (b) Identification of the self-loop. (c) Elimination of the self-loop.

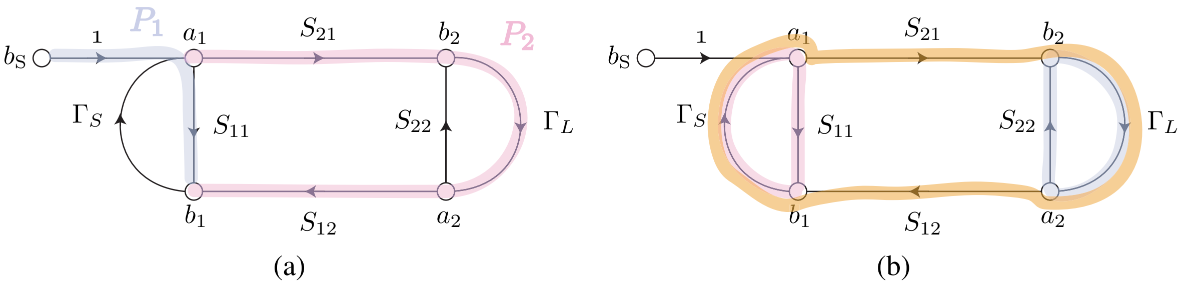


Figure 7.21: (a) Identification of the paths in a signal-flow graph. (b) Identification of the loops in a signal-flow graph.

Example 11:

Input Reflection of Two-Port

Let's redo the calculation of $\Gamma_{in} = b_1/a_1$ for the signal-flow graph shown in Fig. 7.20. Using Mason's rule, you can quickly identify the relevant paths. There are two paths $P_1 = S_{11}$ and $P_2 = S_{21}\Gamma_L S_{12}$. There is only one first-order loop: $\sum \mathscr{L}(1) = S_{22}\Gamma_L$ and so naturally there are no higher order loops. Note that the loop does not touch path P_1, so $\sum \mathscr{L}(1)^{(1)} = S_{22}\Gamma_L$. Now let's apply Mason's general formula

$$\Gamma_{in} = \frac{S_{11}(1 - S_{22}\Gamma_L) + S_{21}\Gamma_L S_{12}}{1 - S_{22}\Gamma_L} = S_{11} + \frac{S_{21}\Gamma_L S_{12}}{1 - S_{22}\Gamma_L}$$

Example 12:

Transducer Power Gain

By definition, the transducer power gain for the two-port shown in Fig. 7.22 is given by

$$G_T = \frac{P_L}{P_{AVS}} = \frac{|b_2|^2(1 - |\Gamma_L|^2)}{\frac{|b_s|^2}{1-|\Gamma_S|^2}}$$

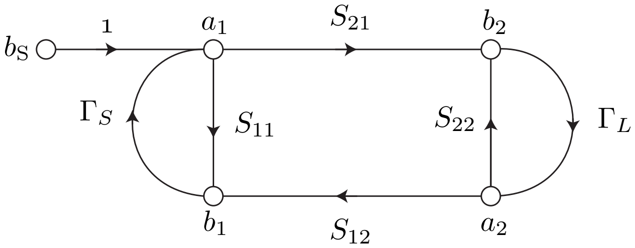


Figure 7.22: A two-port driven by a source with reflection coefficient Γ_S and loaded by Γ_L.

$$= \left|\frac{b_2}{b_S}\right|^2 (1-|\Gamma_L|^2)(1-|\Gamma_S|^2)$$

By Mason's Rule, there is only one path $P_1 = S_{21}$ from b_S to b_2 so we have

$$\sum \mathscr{L}(1) = \Gamma_S S_{11} + S_{22}\Gamma_L + \Gamma_S S_{21}\Gamma_L S_{12}$$

$$\sum \mathscr{L}(2) = \Gamma_S S_{11}\Gamma_L S_{22}$$

$$\sum \mathscr{L}(1)^{(1)} = 0$$

The gain expression is thus given by

$$\frac{b_2}{b_S} = \frac{S_{21}(1-0)}{1-\Gamma_S S_{11} - S_{22}\Gamma_L - \Gamma_S S_{21}\Gamma_L S_{12} + \Gamma_S S_{11}\Gamma_L S_{22}}$$

The denominator is in the form of $1-x-y+xy$ which allows us to write

$$G_T = \frac{|S_{21}|^2(1-|\Gamma_S|^2)(1-|\Gamma_L|^2)}{|(1-S_{11}\Gamma_S)(1-S_{22}\Gamma_L) - S_{21}S_{12}\Gamma_L\Gamma_S|^2}$$

Recall that $\Gamma_{in} = S_{11} + S_{21}S_{12}\Gamma_L/(1-S_{22}\Gamma_L)$. Factoring out $1-S_{22}\Gamma_L$ from the denominator we have

$$\text{den} = \left(1 - S_{11}\Gamma_S - \frac{S_{21}S_{12}\Gamma_L}{1-S_{22}\Gamma_L}\Gamma_S\right)(1-S_{22}\Gamma_L)$$

$$\text{den} = \left(1 - \Gamma_S\left(S_{11} + \frac{S_{21}S_{12}\Gamma_L}{1-S_{22}\Gamma_L}\right)\right)(1-S_{22}\Gamma_L)$$

$$= (1-\Gamma_S\Gamma_{in})(1-S_{22}\Gamma_L)$$

This simplifications allows us to write the transducer gain in the following convenient form

$$G_T = \frac{1-|\Gamma_S|^2}{|1-\Gamma_{in}\Gamma_S|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2}$$

which can be viewed as a product of the action of the input match “gain", the intrinsic two-port gain $|S_{21}|^2$, and the output match “gain". Since the general two-port is not unilateral, the input match is a function of the load. Likewise, by symmetry we can also factor the expression to obtain

$$G_T = \frac{1-|\Gamma_S|^2}{|1-S_{11}\Gamma_S|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-\Gamma_{out}\Gamma_L|^2}$$

7.5 Stability of a Two-Port

A two-port is unstable if the admittance of either port has a negative conductance for a passive termination on the second port. Under such a condition, the two-port can oscillate. Consider the input admittance

$$Y_{in} = G_{in} + jB_{in} = Y_{11} - \frac{Y_{12}Y_{21}}{Y_{22}+Y_L} \tag{7.2}$$

Using the following definitions

$$Y_{11} = g_{11} + jb_{11} \tag{7.3}$$

$$Y_{22} = g_{22} + jb_{22} \tag{7.4}$$

$$Y_{12}Y_{21} = P + jQ = L\angle\phi \tag{7.5}$$

$$Y_L = G_L + jB_L \tag{7.6}$$

Now substitute real/imaginary parts of the above quantities into Y_{in}

$$Y_{in} = g_{11} + jb_{11} - \frac{P+jQ}{g_{22}+jb_{22}+G_L+jB_L} \tag{7.7}$$

$$= g_{11} + jb_{11} - \frac{(P+jQ)(g_{22}+G_L-j(b_{22}+B_L))}{(g_{22}+G_L)^2+(b_{22}+B_L)^2} \tag{7.8}$$

Taking the real part, we have the input conductance

$$\Re(Y_{in}) = G_{in} = g_{11} - \frac{P(g_{22}+G_L)+Q(b_{22}+B_L)}{(g_{22}+G_L)^2+(b_{22}+B_L)^2} \tag{7.9}$$

$$= \frac{(g_{22}+G_L)^2+(b_{22}+B_L)^2 - \frac{P}{g_{11}}(g_{22}+G_L) - \frac{Q}{g_{11}}(b_{22}+B_L)}{D} \tag{7.10}$$

Since $D > 0$ if $g_{11} > 0$, we can focus on the numerator. Note that $g_{11} > 0$ is a requirement since otherwise oscillations would occur for a short circuit at port 2. The numerator can be factored into several positive terms

$$N = (g_{22}+G_L)^2+(b_{22}+B_L)^2 - \frac{P}{g_{11}}(g_{22}+G_L) - \frac{Q}{g_{11}}(b_{22}+B_L) \tag{7.11}$$

$$= \left(G_L + \left(g_{22} - \frac{P}{2g_{11}}\right)\right)^2 + \left(B_L + \left(b_{22} - \frac{Q}{2g_{11}}\right)\right)^2 - \frac{P^2+Q^2}{4g_{11}^2} \tag{7.12}$$

Now note that the numerator can go negative only if the first two terms are smaller than the last term. To minimize the first two terms, choose $G_L = 0$ and $B_L = -\left(b_{22} - \frac{Q}{2g_{11}}\right)$ (reactive load)

$$N_{min} = \left(g_{22} - \frac{P}{2g_{11}}\right)^2 - \frac{P^2 + Q^2}{4g_{11}^2} \tag{7.13}$$

And thus the above must remain positive, $N_{min} > 0$, so

$$\left(g_{22} - \frac{P}{2g_{11}}\right)^2 - \frac{P^2 + Q^2}{4g_{11}^2} > 0 \tag{7.14}$$

$$g_{11}g_{22} > \frac{P + L}{2} = \frac{L}{2}(1 + \cos\phi) \tag{7.15}$$

Linvill/Llewellyn Stability Factors

Using the above equation, we define the Linvill stability factor

$$L < 2g_{11}g_{22} - P \tag{7.16}$$

$$C = \frac{L}{2g_{11}g_{22} - P} < 1 \tag{7.17}$$

The two-port is stable if $0 < C < 1$. It's more common to use the inverse of C as the stability measure

$$\frac{2g_{11}g_{22} - P}{L} > 1 \tag{7.18}$$

The above definition of stability is perhaps the most common

$$K = \frac{2\Re(Y_{11})\Re(Y_{22}) - \Re(Y_{12}Y_{21})}{|Y_{12}Y_{21}|} > 1 \tag{7.19}$$

The above expression is identical if we interchange ports 1/2. Thus it's the general condition for stability. Note that $K > 1$ is the same condition for the maximum stable gain derived last section. The connection is now more obvious. If $K < 1$, then the maximum gain is infinity!

7.5.1 Stability from Scattering Parameters

We can also derive stability in terms of the input reflection coefficient. For a general two-port with load Γ_L we have

$$v_2^- = \Gamma_L^{-1} v_2^+ = S_{21}v_1^+ + S_{22}v_2^+ \tag{7.20}$$

$$v_2^+ = \frac{S_{21}}{\Gamma_L^{-1} - S_{22}} v_1^- \tag{7.21}$$

$$v_1^- = \left(S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - \Gamma_L S_{22}}\right) v_1^+ \tag{7.22}$$

$$\Gamma = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1-\Gamma_L S_{22}} \tag{7.23}$$

If $|\Gamma| < 1$ for all Γ_L, then the two-port is stable

$$\Gamma = \frac{S_{11}(1-S_{22}\Gamma_L)+S_{12}S_{21}\Gamma_L}{1-S_{22}\Gamma_L} = \frac{S_{11}+\Gamma_L(S_{21}S_{12}-S_{11}S_{22})}{1-S_{22}\Gamma_L} \tag{7.24}$$

$$= \frac{S_{11}-\Delta\Gamma_L}{1-S_{22}\Gamma_L} \tag{7.25}$$

To find the boundary between stability/instability, let's set $|\Gamma| = 1$

$$\left|\frac{S_{11}-\Delta\Gamma_L}{1-S_{22}\Gamma_L}\right| = 1 \tag{7.26}$$

$$|S_{11}-\Delta\Gamma_L| = |1-S_{22}\Gamma_L| \tag{7.27}$$

After some algebraic manipulations, we arrive at the following equation

$$\left|\Gamma - \frac{S_{22}^*-\Delta^* S_{11}}{|S_{22}|^2-|\Delta|^2}\right| = \frac{|S_{12}S_{21}|}{|S_{22}|^2-|\Delta|^2} \tag{7.28}$$

This is of course the equation of a circle, $|\Gamma - C| = R$, in the complex plane with center at C and radius R. Thus a circle on the Smith Chart divides the region of instability from stability.

Consider the stability circle for a unilateral two-port

$$C_S = \frac{S_{11}^*-(S_{11}^*S_{22}^*)S_{22}}{|S_{11}|^2-|S_{11}S_{22}|^2} = \frac{S_{11}^*}{|S_{11}|^2} \tag{7.29}$$

$$R_S = 0 \tag{7.30}$$

$$|C_S| = \frac{1}{|S_{11}|} \tag{7.31}$$

The center of the circle lies outside of the unit circle if $|S_{11}| < 1$. The same is true of the load stability circle. Since the radius is zero, stability is only determined by the location of the center. If $S_{12} = 0$, then the two-port is unconditionally stable if $S_{11} < 1$ and $S_{22} < 1$. This result is trivial since

$$\Gamma_S|_{S_{12}=0} = S_{11} \tag{7.32}$$

The stability of the source depends only on the device and not on the load.

7.5.2 μ Stability Test

If we want to determine if a two-port is unconditionally stable, then we should use the μ-test

$$\mu = \frac{1-|S_{11}|^2}{|S_{22}-\Delta S_{11}^*|+|S_{12}S_{21}|} > 1 \tag{7.33}$$

The μ-test not only is a test for unconditional stability, but the magnitude of μ is a measure of the stability. In other words, if one two-port has a larger μ, it is more stable.

The advantage of the μ-test is that only a single parameter needs to be evaluated. There are no auxiliary conditions like the K-test derivation earlier. The derivation of the μ-test can proceed as follows. First let $\Gamma_S = |\rho_s|e^{j\phi}$ and evaluate Γ_{out}

$$\Gamma_{out} = \frac{S_{22}-\Delta|\rho_s|e^{j\phi}}{1-S_{11}|\rho_s|e^{j\phi}} \tag{7.34}$$

Next we can manipulate this equation into the equation for a circle $|\Gamma_{out}-C| = R$

$$\left|\Gamma_{out} + \frac{|\rho_s|S_{11}^*\Delta - S_{22}}{1-|\rho_s||S_{11}|^2}\right| = \frac{\sqrt{|\rho_s|}|S_{12}S_{21}|}{(1-|\rho_s||S_{11}|^2)} \tag{7.35}$$

For a two-port to be unconditionally stable, we'd like Γ_{out} to fall within the unit circle

$$||C|+R| < 1 \tag{7.36}$$

$$||\rho_s|S_{11}^*\Delta - S_{22}| + \sqrt{|\rho_s|}|S_{21}S_{12}| < 1-|\rho_s||S_{11}|^2 \tag{7.37}$$

$$||\rho_s|S_{11}^*\Delta - S_{22}| + \sqrt{|\rho_s|}|S_{21}S_{12}| + |\rho_s||S_{11}|^2 < 1 \tag{7.38}$$

The worst case stability occurs when $|\rho_s| = 1$ since it maximizes the left-hand side of the equation. Therefore we have

$$\mu = \frac{1-|S_{11}|^2}{|S_{11}^*\Delta - S_{22}|+|S_{12}S_{21}|} > 1 \tag{7.39}$$

K-Δ Test

The K stability test has already been derived using Y parameters. We can also do a derivation based on S parameters. This form of the equation has been attributed to Rollett and Kurokawa. The idea is very simple and similar to the μ test. We simply require that all points in the instability region fall outside of the unit circle. The stability circle will intersect with the unit circle if

$$|C_L| - R_L > 1 \tag{7.40}$$

or

$$\frac{|S_{22}^* - \Delta^* S_{11}| - |S_{12}S_{21}|}{|S_{22}|^2 - |\Delta|^2} > 1 \tag{7.41}$$

This can be recast into the following form (assuming $|\Delta| < 1$)

$$K = \frac{1-|S_{11}|^2-|S_{22}|^2+|\Delta|^2}{2|S_{12}||S_{21}|} > 1 \tag{7.42}$$

7.5.3 N-Port Passivity

We would like to find if an N-port is active or passive. Passivity is different from stability, and plays an important role in determining the maximum frequency of operation for an "active" device. For instance, above a certain frequency every transistor will transition from an active device to a passive device, setting an upper limit for amplification or oscillation with a given device. By definition, an N-port is passive if it can only absorb net power. The total net complex power flowing into or out of a N port is given by

$$P = (V_1^* I_1 + V_2^* I_2 + \cdots) = (I_1^* V_1 + I_2^* V_2 + \cdots) \tag{7.43}$$

If we sum the above two terms we have

$$P = \frac{1}{2}(v^*)^T i + \frac{1}{2}(i^*)^T v \tag{7.44}$$

for vectors of current and voltage i and v. Using the admittance matrix $i = Yv$, this can be recast as

$$P = \frac{1}{2}(v^*)^T Y v + \frac{1}{2}(Y^* v^*)^T v = \frac{1}{2}(v^*)^T Y v + \frac{1}{2}(v^*)^T (Y^*)^T v \tag{7.45}$$

$$P = (v^*)^T \frac{1}{2}(Y + (Y^*)^T)v = (v^*)^T Y_H v \tag{7.46}$$

Thus for a network to be passive, the Hermitian part of the matrix Y_H should be positive semi-definite.

For a two-port, the condition for passivity can be simplified as follows. Let the general hybrid admittance matrix for the two-port be given by

$$H(s) = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} + j \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix} \tag{7.47}$$

$$H_H(s) = \frac{1}{2}(H(s) + H^*(s)) \tag{7.48}$$

$$= \begin{pmatrix} m_{11} & \frac{1}{2}((m_{12} + m_{21}) + j(n_{12} - n_{21})) \\ ((m_{12} + m_{21}) + j(n_{21} - n_{12})) & m_{22} \end{pmatrix} \tag{7.49}$$

This matrix is positive semi-definite if

$$m_{11} > 0 \tag{7.50}$$

$$m_{22} > 0 \tag{7.51}$$

$$\det H_n(s) \geq 0 \tag{7.52}$$

or

$$4 m_{11} m_{22} - |k_{12}|^2 - |k_{21}|^2 - 2\Re(k_{12} k_{21}) \geq 0 \tag{7.53}$$

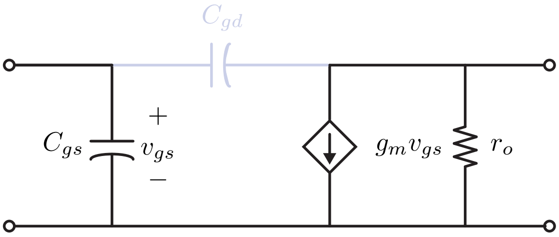


Figure 7.23: A simplified hybrid-π equivalent circuit.

$$4m_{11}m_{22} \geq |k_{12}+k_{21}^*|^2 \tag{7.54}$$

Example 13:

A simple equivalent circuit for a FET without any feedback, shown in Fig. 7.23, is of course absolutely stable if the resistors of the model are positive. The Z matrix for the circuit is given by

$$Z = \begin{bmatrix} \frac{1}{j\omega C_{gs}} & 0 \\ \frac{-g_m r_o}{j\omega C_{gs}} & r_o \end{bmatrix} \tag{7.55}$$

Since $Z_{12} = 0$, the stability factor $K = \infty$

$$K = \frac{2\Re(Z_{11})\Re(Z_{22}) - \Re(Z_{12}Z_{21})}{|Z_{12}Z_{21}|} \tag{7.56}$$

Example 14:

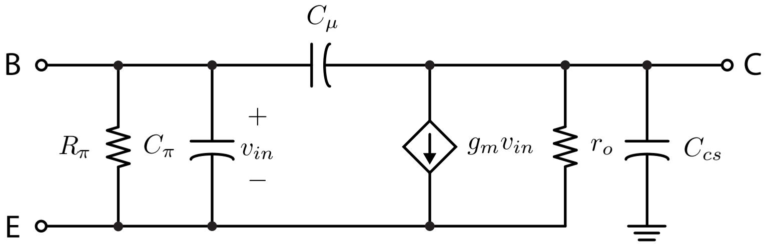


Figure 7.24: The simple hybrid-pi model for a transistor.

The hybrid-pi model for a transistor is shown in Fig. 7.24. Under what conditions is this two-port active? The hybrid matrix is given by

$$H(s) = \frac{1}{G_\pi + s(C_\pi + C_\mu)} \begin{pmatrix} 1 & sC_\mu \\ g_m - sC_\mu & q(s) \end{pmatrix} \tag{7.57}$$

$$q(s) = (G_\pi + sC_\pi)(G_0 + sC_\mu) + sC_\mu(G_\pi + g_m) \tag{7.58}$$

Applying the condition for passivity we arrive at

$$4G_\pi G_0 \geq g_m^2 \tag{7.59}$$

The above equation is either satisfied for the two-port or not, regardless of frequency. Thus our analysis shows that the hybrid-pi model is not physical. We know from experience that real two-ports are active up to some frequency f_{max}.

7.5.4 Mason's Invariant U Function

In 1954, Samuel Mason discovered the function U given by [**mason**]

$$U = \frac{|k_{21} - k_{12}|^2}{4(\Re(k_{11})\Re(k_{22}) - \Re(k_{12})\Re(k_{21}))} \tag{7.60}$$

For the hybrid matrix formulation (H or G), the U function is given by

$$U = \frac{|k_{21} + k_{12}|^2}{4(\Re(k_{11})\Re(k_{22}) + \Re(k_{12})\Re(k_{21}))} \tag{7.61}$$

where k_{ij} are the two-port Y, Z, H, or G parameters.

This function is invariant under lossless reciprocal embeddings. Stated differently, any two-port can be *embedded* into a lossless and reciprocal circuit and the resulting two-port will have the same U function. This is a very important property, because this invariant property does not depend on any lossless matching circuitry that we employ before or after the two-port, or any lossless feedback.

Properties of U

The invariant property is shown in Fig. 7.25. The U of the original two-port is the same as U_a of the overall two-port when a four port lossless reciprocal four-port is added.

The U function has several important properties:

1. If $U > 1$, the two-port is active. Otherwise, if $U \leq 1$, the two-port is passive.
2. U is the maximum unilateral power gain of a device under a lossless reciprocal embedding.
3. U is the maximum gain of a three-terminal device regardless of the common terminal.

With regards to the previous diagram, any lossless reciprocal embedding can be seen as an interconnection of the original two-port to a four-port, with the following block admittance matrix [**waikaichen**]

$$\begin{pmatrix} I_a \\ -I \end{pmatrix} = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 \\ Y_{21}^0 & Y_{22}^0 \end{pmatrix} \begin{pmatrix} V_a \\ V \end{pmatrix} \tag{7.62}$$

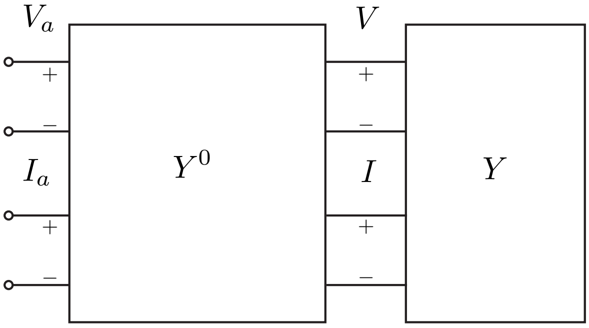


Figure 7.25: A general two-port described by Y is embedded into a lossless, reciprocal four-port device described by the matrix Y^0.

Note that Y_{ij} is a 2×2 imaginary symmetric sub-matrix

$$Y^0_{jk} = jB_{jk} \tag{7.63}$$

$$B_{jk} = B^T_{kj} \tag{7.64}$$

Since $I = YV$, we can solve for V from the second equation

$$-I = Y^0_{21}V_a + Y^0_{22}V = -YV \tag{7.65}$$

$$V = -(Y + Y^0_{22})^{-1}Y^0_{21}V_a \tag{7.66}$$

From the first equation we have the composite two-port matrix

$$I_a = (Y^0_{11} - Y^0_{12}(Y + Y^0_{22})^{-1}Y^0_{21})V_a = Y_aV_a \tag{7.67}$$

By definition, the U function is given by

$$U = \frac{\det(Y_a - Y_a^T)}{\det(Y_a + Y_a^*)} \tag{7.68}$$

Note that Y_a can be written as

$$Y_a = jB_{11} - jB_{12}(Y + jB_{22})^{-1}jB^T_{12} \tag{7.69}$$

$$Y_a = jB_{11} + B_{12}(Y + jB_{22})^{-1}B^T_{12} \tag{7.70}$$

Focus on the denominator of U

$$Y_a + Y_a^* = B_{12}(W^{-1} + (W^*)^{-1})B^T_{12} \tag{7.71}$$

where $W = Y + Y^0_{22} = Y + jB_{22}$. Factoring W^{-1} from the left and $(W^*)^{-1}$ from the right, we have

$$= B_{12}W^{-1}(W^* + W)(W^*)^{-1}B^T_{12} \tag{7.72}$$

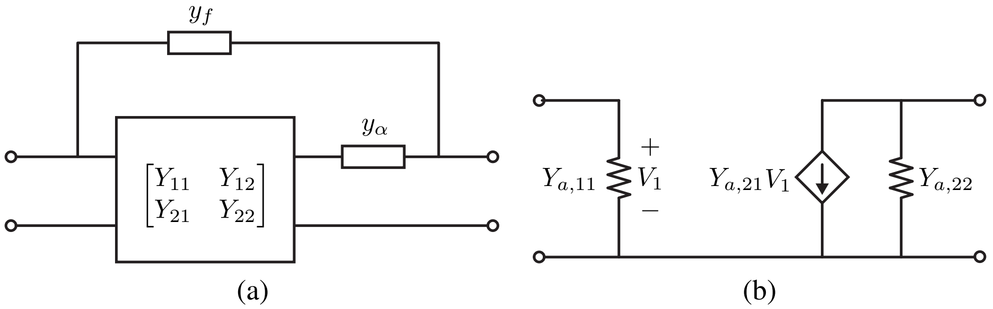

y_f
y_α
$\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$
$Y_{a,11}$
$+$
V_1
$-$
$Y_{a,21}V_1$
$Y_{a,22}$
(a)
(b)

Figure 7.26: (a) A general two-port can be *unilaterized* by adding lossless feedback elements y_f and y_α. (b) The equivalent circuit for the unilaterized two-port.

But $W + W^* = Y + Y^*$ resulting in

$$Y_a + Y_a^* = B_{12} W^{-1} (Y + Y^*)(W^*)^{-1} B_{12}^T \tag{7.73}$$

In a like manner, one can show that

$$Y_a - Y_a^T = B_{12} W^{-1} (Y^T - Y)(W^*)^{-1} B_{12}^T \tag{7.74}$$

Taking the determinants and ratios

$$\det(Y_a + Y_a^*) = \frac{(\det B_{12})^2 \det(Y + Y^*)}{(\det W)^2} \tag{7.75}$$

$$\det(Y_a - Y_a^T) = \frac{(\det B_{12})^2 \det(Y^T - Y)}{(\det W)^2} \tag{7.76}$$

$$U = \frac{\det(Y_a - Y_a^T)}{\det(Y_a + Y_a^*)} = \frac{\det(Y - Y^T)}{\det(Y + Y^*)} \tag{7.77}$$

Maximum Unilateral Gain

Consider Fig. 7.26a, a feedback structure where y_f and y_α are lossless reactances. We can derive the overall two-port equations by a cascade connection followed by a shunt connection of two-ports

$$Y_a = \frac{y_\alpha}{y_\alpha + y_{22}} \begin{bmatrix} y_{11} + \Delta_y / y_\alpha & y_{12} \\ y_{21} & y_{22} \end{bmatrix} + \begin{bmatrix} y_f & -y_f \\ -y_f & y_f \end{bmatrix} \tag{7.78}$$

To unilaterize the device, we select

$$y_f = \frac{y_{12} y_\alpha}{y_{22} + y_\alpha} \tag{7.79}$$

We can solve for b_α and b_f

$$b_f = \Im(y_{12}) - \frac{\Re(y_{12})}{\Re(y_{22})} \Im(y_{22}) \tag{7.80}$$

$$b_\alpha = b_f \frac{\Re(y_{22})}{\Re(y_{12})} \tag{7.81}$$

It can be shown that the overall Y_a matrix is given by

$$Y_a = \frac{j\Im(y_{22}^* y_{12})}{y_{12} \Re(y_{22})} \begin{bmatrix} y_{11} + y_{12} - j\frac{\Delta_y \Re(y_{12})}{\Im(y_{22}^* y_{12})} & 0 \\ y_{21} - y_{12} & y_{22} + y_{12} \end{bmatrix} \tag{7.82}$$

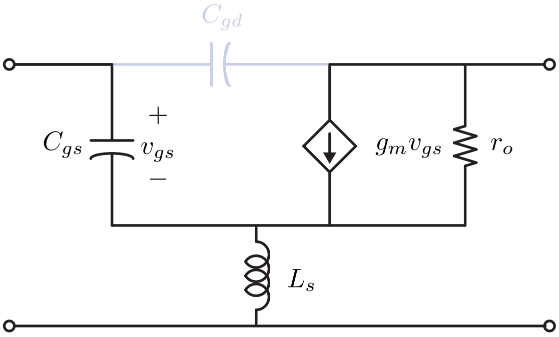


Figure 7.27: A FET with inductive degeneration.

Unilaterized Two-Port

The two-port equivalent circuit under unilaterization is shown in Fig. 7.26b. Notice now that the maximum power gain of this circuit is given by

$$G_{U,max} = \frac{|Y_{a_{21}}|^2}{4\Re(Y_{a_{11}})\Re(Y_{a_{22}})} = U_a \tag{7.83}$$

We can now attribute physical significance to U_a as the maximum unilateral gain. Furthermore, due to the invariance of U, $U_a = U$ for the original two-port network. It's important to note that *any* unilaterization scheme will yield the same maximum power! Thus U is a good metric for the device.

Single Stage Feedback Revisited

With new tools at hand, let's revisit the problem of inductive and shunt feedback amplifiers.

Inductive Degeneration

Although $Z_{12} \approx 0$ for a FET at low frequency, the input impedance is purely capacitive. To introduce a real component, we found that inductive degeneration can be employed, shown schematically in Fig. 7.27. The Z matrix for the inductor is simply

$$Z_L = j\omega L_s \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{7.84}$$

Adding the Z matrix (due to series connection) to the Z matrix of the FET

$$Z = \begin{bmatrix} j\omega L_s + \frac{1}{j\omega C_{gs}} & j\omega L_s \\ j\omega L_s - \frac{g_m r_o}{j\omega C_{gs}} & r_o + j\omega L_s \end{bmatrix} \tag{7.85}$$

This feedback introduces a Z_{12} and thus the stability must be carefully examined

$$K = \frac{2 \cdot 0 \cdot r_o - \left(-\omega^2 L_s^2 - \frac{g_m L_s r_o}{C_{gs}}\right)}{\omega^2 L_s^2 + \frac{g_m r_o L_s}{C_{gs}}} = 1 \tag{7.86}$$

We see that this circuit is unconditionally stable. More importantly, the stability factor is frequency independent. In reality parasitics can destabilize the transistor.

The maximum gain is thus given by

$$G_{max} = \left|\frac{Z_{21}}{Z_{12}}\right| \left(K - \sqrt{K^2 - 1}\right) = \left|\frac{Z_{21}}{Z_{12}}\right| \tag{7.87}$$

$$= \frac{\omega L_s + \frac{g_m r_o}{\omega C_{gs}}}{\omega L_s} = 1 + \frac{g_m r_o}{\omega^2 L_s C_{gs}} \tag{7.88}$$

$$= 1 + \left(\frac{\omega_T}{\omega_0}\right)^2 \left(\frac{r_o}{\omega_T L_s}\right) \tag{7.89}$$

The synthesized real input resistance is given by $\omega_T L_s$, and so the last term is the ratio of r_o/R_S under matched conditions.

Capacitive Degeneration

A capacitively degenerated transistor is an important building block for Colpitts oscillators, where instability is desired. Using the same approach, the Z matrix for capacitive degeneration is given by

$$Z = \begin{bmatrix} \frac{1}{j\omega C_s} + \frac{1}{j\omega C_{gs}} & \frac{1}{j\omega C_s} \\ \frac{1}{j\omega C_s} - \frac{g_m r_o}{j\omega C_{gs}} & r_o + \frac{1}{j\omega C_s} \end{bmatrix} \tag{7.90}$$

The stability factor is given by

$$K = \frac{2 \cdot 0 \cdot r_o - \left(\frac{g_m r_o}{\omega^2 C_s C_{gs}} - \frac{1}{\omega^2 C_s^2}\right)}{\left|\frac{g_m r_o}{\omega^2 C_s C_{gs}} - \frac{1}{\omega^2 C_s^2}\right|} \tag{7.91}$$

Note this is simply

$$K = \frac{-a+b}{|a-b|} = \begin{cases} \frac{b-a}{a-b} < 0 & a > b \\ \frac{b-a}{b-a} = 1 & b < a \end{cases} \tag{7.92}$$

The condition for stability is therefore

$$\frac{g_m r_o}{C_{gs}} > \frac{1}{C_s} \tag{7.93}$$

So far we have dealt with $K > 0$. Suppose that $|\Delta| > 1$. We know that for $0 < K < 1$ the two-port is conditionally stable. In other words, the stability circle intersects with the unit circle with the overlap (usually) corresponding to the unstable region. Instability can also occur if $K > 1$ and $|\Delta| > 1$, but this is less common (occurs with feedback).

On the other hand, if $-1 < K < 0$, one can show graphically that the entire unit circle on the Smith Chart is unstable. In other words, the stability circle does not intersect with the unit circle or the instability circle contains the entire circle.

Unintentional capacitive degeneration is very common. For instance a common drain (source follower) driving a capacitive load may have stability problems. Likewise, a cascode amplifier may become unstable at high frequencies since the g_m input stage presents capacitive degeneration to the cascode device at high frequency.

Resistive Degeneration

Resistive degeneration is commonly employed to stabilize the bias point of a transistor. The overall Z matrix is given by

$$Z = \begin{bmatrix} R_s + \frac{1}{j\omega C_{gs}} & R_s \\ R_s - \frac{g_m r_o}{j\omega C_{gs}} & r_o + R_s \end{bmatrix} \tag{7.94}$$

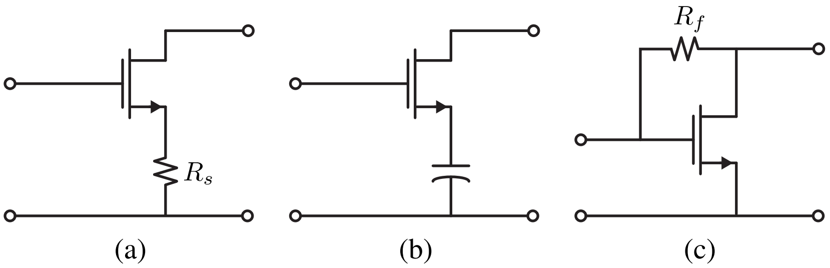


Figure 7.28: Common source amplifier with (a) capacitive degeneration, (b) resistive degeneration, (c) shunt feedback.

The K factor is computed as before

$$K = \frac{2R_s(r_o + R_s) - R_s^2}{R_s\sqrt{R_s^2 + \frac{g_m^2 r_o^2}{\omega^2 C_{gs}^2}}} \tag{7.95}$$

At low frequencies, we have

$$K = \frac{2r_o + R_s}{\frac{g_m r_o}{\omega C_{gs}}} \approx \frac{2\omega C_{gs}}{g_m} = \frac{2\omega}{\omega_T} < 1 \tag{7.96}$$

Shunt Feedback

We have seen that shunt feedback is a common broadband matching approach. Now working with the Y matrix of the transistor (simplified as before)

$$Y_{fet} = \begin{bmatrix} j\omega C_{gs} & 0 \\ g_m & G_o + j\omega C_{ds} \end{bmatrix} \tag{7.97}$$

The feedback element has a Y matrix

$$Y_f = G_f \begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix} \tag{7.98}$$

And thus the overall amplifier Y matrix is given by

$$Y = \begin{bmatrix} G_f + j\omega C_{gs} & -G_f \\ g_m - G_f & G_f + G_o + j\omega C_{ds} \end{bmatrix} \tag{7.99}$$

The stability factor for the shunt feedback amplifier is given by

$$K = \frac{2G_f(G_o + G_f) - G_f(G_f - g_m)}{G_f|g_m - G_f|} \tag{7.100}$$

Suppose that $g_m R_f > 1$

$$= \frac{g_m + G_f}{g_m - G_f} = \frac{g_m R_f + 1}{g_m R_f - 1} > 1 \tag{7.101}$$

The choice of R_f and g_m is governed by the current consumption, power gain, and impedance matching. For a bi-conjugate match

$$G_{max} = \left|\frac{Y_{21}}{Y_{12}}\right| \left(K - \sqrt{K^2 - 1}\right) \tag{7.102}$$

$$= \frac{g_m - G_f}{G_f}\left(\left(\frac{g_m R_f + 1}{g_m R_f - 1}\right) - \sqrt{\left(\frac{g_m R_f + 1}{g_m R_f - 1}\right)^2 - 1}\right) = \left(1 - \sqrt{g_m R_F}\right)^2 \tag{7.103}$$

The input admittance is calculated as follows

$$Y_{in} = Y_{11} - \frac{Y_{12}Y_{21}}{Y_{22} + Y_L} \tag{7.104}$$

$$= j\omega C_{gs} + G_f - \frac{-G_f(g_m - G_f)}{G_o + G_f + G_L + j\omega C_{ds}} \tag{7.105}$$

$$= j\omega C_{gs} + G_f + \frac{G_f(g_m - G_f)(G_o + G_f + G_L - j\omega C_{ds})}{(G_o + G_f + G_L)^2 + \omega^2 C_{ds}^2} \tag{7.106}$$

At lower frequencies, $\omega < \frac{1}{C_{ds} R_f || R_L}$ we have (neglecting G_o)

$$\Re(Y_{in}) = G_f + \frac{G_f(g_m - G_f)}{G_f + G_L} \tag{7.107}$$

$$= \frac{1 + g_m R_L}{R_F + R_L} \tag{7.108}$$

$$\Im(Y_{in}) = \omega\left(C_{gs} - \frac{C_{ds}}{1 + \frac{R_f}{R_L}}\right) \tag{7.109}$$

7.6 Transistor Figures of Merit

A common figure of merit to characterize transistors is the device unity gain frequency, f_T, which are connected to the fundamental device physics. But RF device characterization is based upon f_{max}, or the maximum frequency where we can extract power gain from the device. Essentially, beyond the f_{max} frequency, the device is passive and it cannot be used to build an amplifier with power gain. Likewise, beyond f_{max} one cannot build an oscillator from an amplifier since oscillators need nearly infinite power gain, usually realized through feedback.[1] If a device does not have power gain, it certainly cannot have infinite power gain with feedback, and so the f_{max} frequency also corresponds to the maximum frequency of oscillation.

By definition, therefore, the frequency point where G_{max} crosses unity is the f_{max} of a two-port. Recall that G_{max} is only defined when the transistor is unconditionally stable, or $K > 1$. If $K < 1$, G_{max} is undefined and we usually speak of the maximum stable gain G_{MSG}, which corresponds to the maximum gain when the transistor is stabilized by adding positive conductance at the input and/or output ports so that $K = 1$.

In practice, the device f_{max} is usually estimated by plotting the device maximum unilateral power gain, or Mason's Gain U, and either observing or extrapolating the unity gain frequency point. This procedure should be performed with care and extrapolations should be avoided for maximal accuracy. If data is not available (e.g. above 100 GHz), it's better to model the device

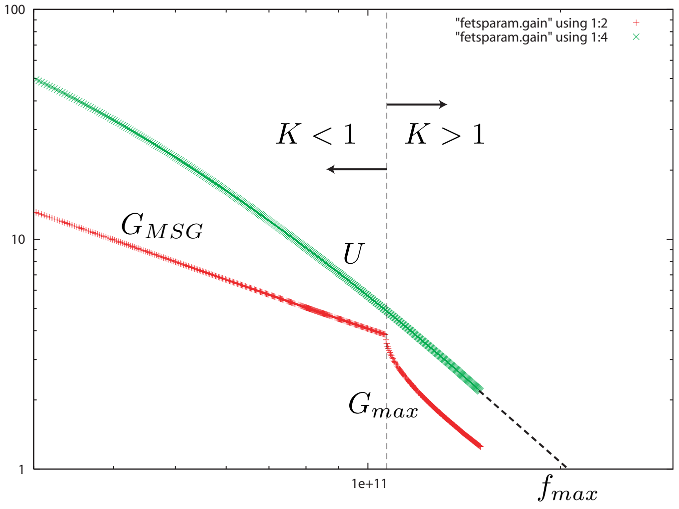


Figure 7.29: The various gain curves for a two-port device. The device is unstable at low frequency, $K < 1$, and thus we plot the G_{MSG} in this region. At the breakpoint, the device is stable. At high frequency the device is stable and we plot the G_{max} curve. The maximum unilateral gain U is also shown.

with an equivalent circuit up to the limits of measurements and then to use the f_{max} from directly evaluating the model up to the point when the power gain crosses unity.

In Fig. 7.29, the device G_{MSG} is plotted for low frequencies where $K < 1$. At the breakpoint, $K = 1$ and the device is unconditionally stable and thus G_{max} is plotted. Note that the U curve is always larger than G_{max} but both curves cross 0 dB together. At this point, the f_{max} of the device, the two-port becomes passive. f_{max} is a good metric for characterizing a three terminal device with a common-terminal, such as a transistor. Since U is invariant to the common terminal, a common-gate amplifier has the same U as a common-source amplifier.

Using the unilateral gain U, the f_{max} of a BJT transistor can be estimated by

$$f_{max} \approx \sqrt{\frac{f_T}{8\pi r_b C_\mu}} \tag{7.110}$$

where the base resistance r_b and feedback capacitance C_μ are seen to set the ultimate frequency of operation for a device. It's interesting to observe that in most low frequency design, both of these effects are ignored with negligible error. But design close to the limits of a device f_{max} requires careful modeling of all parasitic feedback and loss mechanisms. In particular, the distributed nature of the feedback C_μ into r_b requires a sectional model.

The cross section of a MOSFET device is shown in Fig. 7.30. The f_{max} of a modern FET transistor can be estimated by [**wcm03**]

$$f_{max} \approx \frac{f_T}{2\sqrt{R_g\left(g_m C_{gd}/C_{gg}\right) + \left(R_g + r_{ch} + R_S\right) g_{ds}}} \tag{7.111}$$

In contrast to the device f_T, the f_{max} is a strong function of the losses in the device. As MOS technology scaling continues, f_T improves almost proportional to channel length due to velocity saturation. But shorter channel devices may have higher gate, source and drain losses. It is

[1] Nearly infinite because in any real circuit there is noise and thus the oscillator power gain is extremely large, but not infinite.

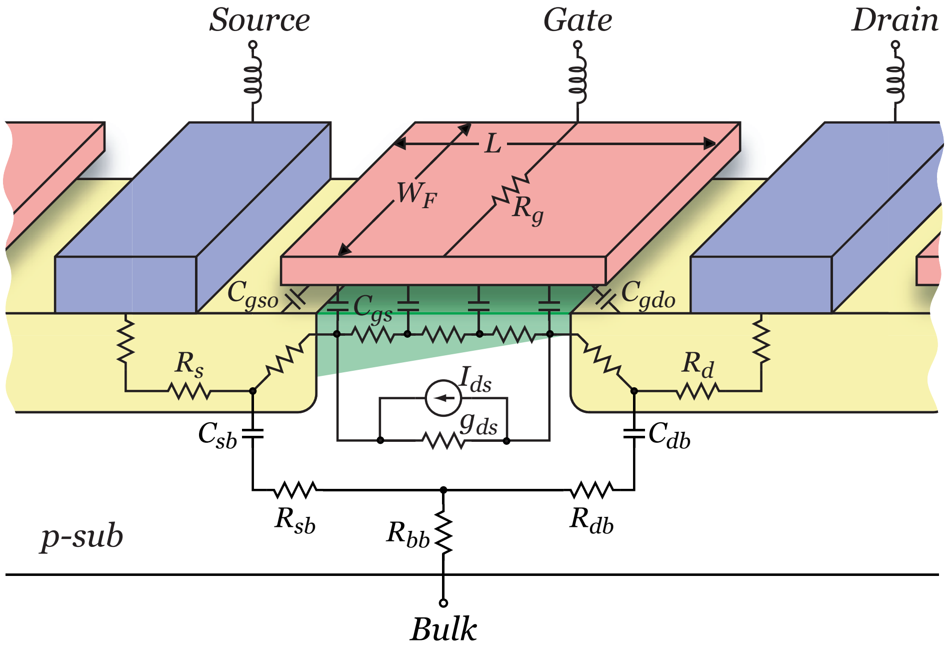


Figure 7.30: The cross section of a FET device showing the important high frequency parasitics. *Courtesy of Chinh Doan.*

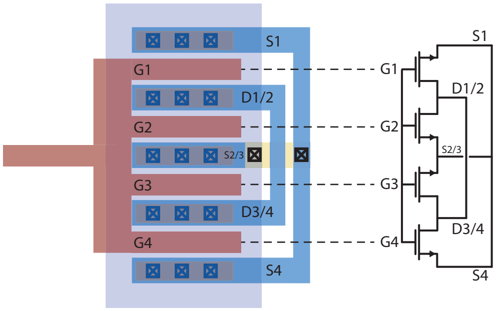
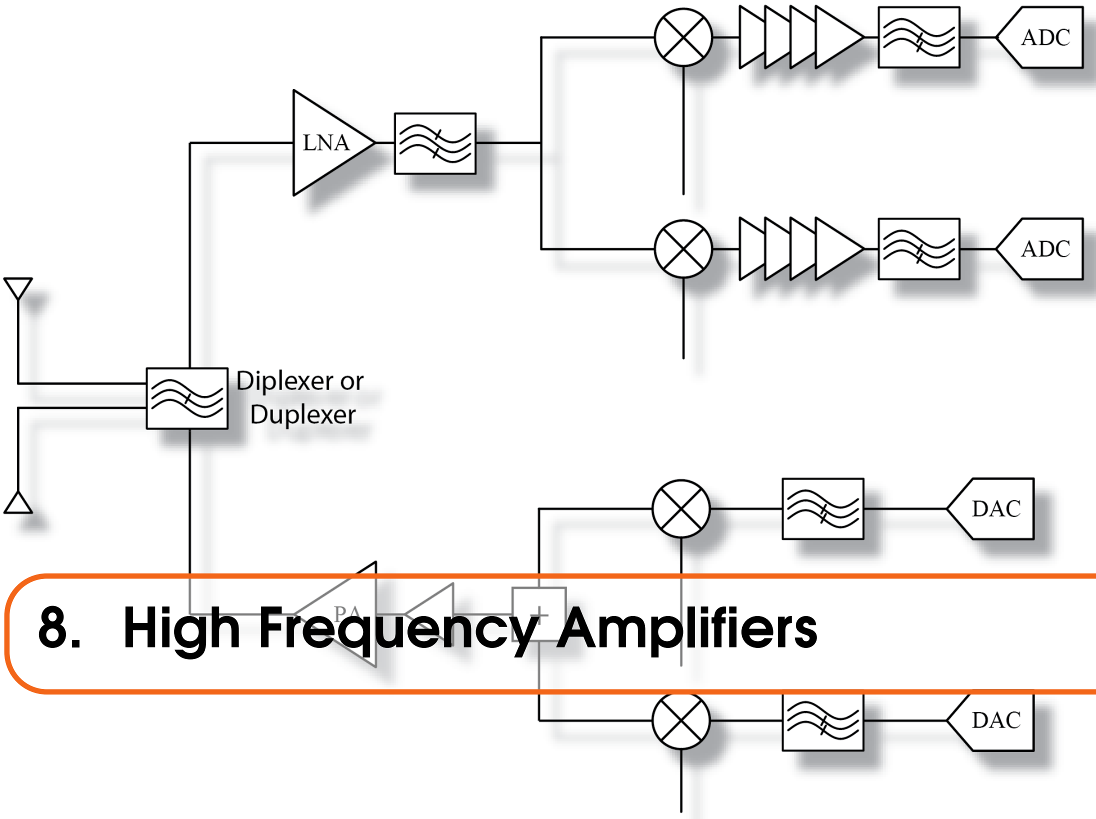


Figure 7.31: A high frequency multi-finger FET layout minimizes the poly gate resistance.

interesting to note that the effect of gate resistance on f_{max} can be reduced by scaling the width of the transistor W through a multi-finger layout, as shown in Fig. 11.12. The drain and source resistances, though, do not scale and pose a challenge for next-generation technologies. This is in stark contrast to MESFET devices where a low resistance metal gate is employed. In deeply scaled MOS technology, metal gate work-function engineering may replace doping as a means to set the threshold voltage of a device, leading to enhanced RF performance.



8. High Frequency Amplifiers

8.1 Introduction

Amplifiers are key building blocks in any communication system. In a receiver, the weak incoming signal needs to be amplified to a sufficiently large value so that it can be detected or digitized. On the transmit side, the signal amplitude needs to be large enough for long-range transmission through free space or cables. In this chapter we will address the design high-frequency narrowband and broadband amplifier.

8.2 MOS Technology

The cross section of a modern CMOS process is shown in Fig. 8.1. The figure shows a p-type substrate with a substrate NMOS and an n-well PMOS device. In a triple well process, isolated NMOS devices can also be fabricated in an n-well. Modern short channel CMOS process has

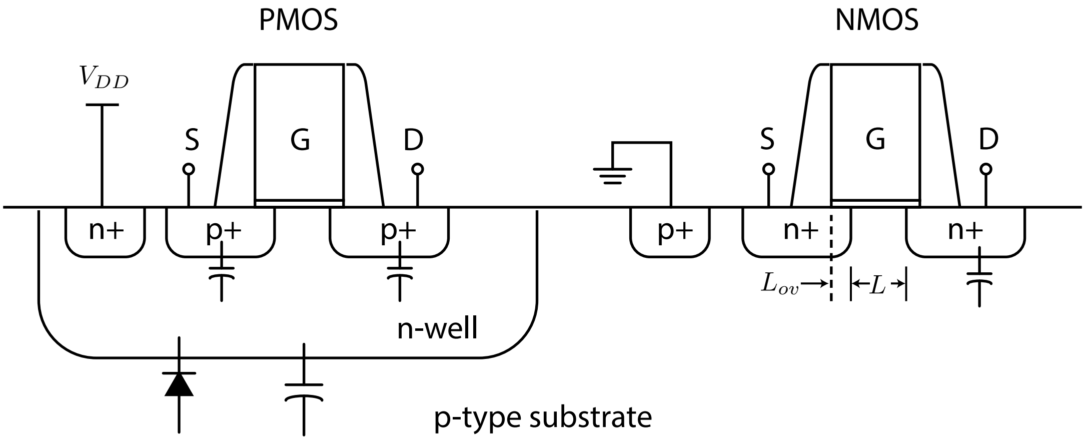


Figure 8.1: The cross-section of a modern CMOS process.

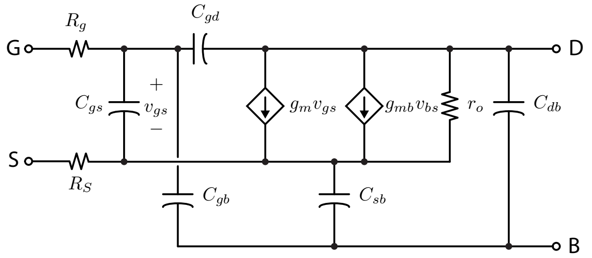


Figure 8.2: The small-signal equivalent circuit model of a MOSFET device.

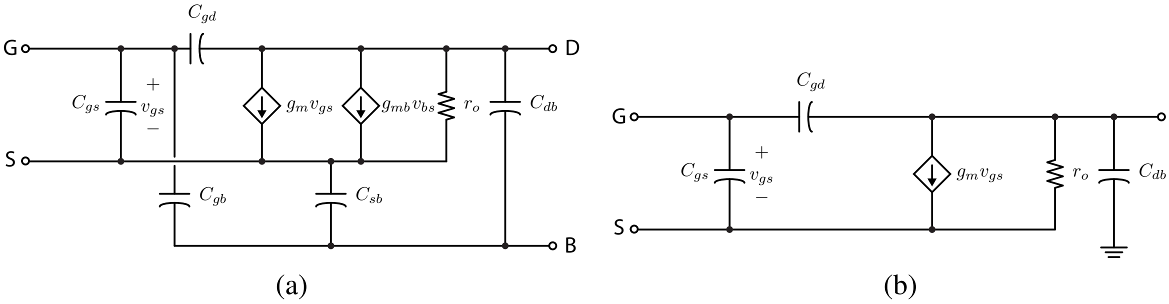


Figure 8.3: (a) The complete and (b) simplified small-signal model of a FET in saturation.

very short channel lengths ($L < 10\,\text{nm}$). To ensure gate control of channel, as opposed to drain control (DIBL), drain and source junctions are fabricated to be as thin as possible while the oxide thikness is reduced to quantum (leakage due to tunneling) limits ($t_{ox} < 5\,\text{nm}$). Due to lithographic limitations, there is an overlap between the gate and the source/drain junctions, which leads to overlap capacitance ($C_{gs,ov}$,$C_{gd,ov}$). In a modern FET this is a substantial fraction of the gate capacitance C_{gs}.

In circuit design the transistor is often modeled using equivalent circuits, which allow one to understand the functionality of a transistor in a simple compact and graphical manner. We arrive at the small-signal model through linearizing the large signal equations at a particular fixed (static) operating point.

The FET small-signal model shown is in Fig. 8.2. The junctions of a FET form reverse-biased pn junctions with the substrate (well), or the body node. This is a form of parasitic capacitance in the structure, represented by C_{db} and C_{sb}. At low frequencies, the gate terminal is nearly open and $R_{in} \sim \infty$. At intermediate frequencies, the input impedance is dominated by C_{gs}, although there is also external gate resistance R_g due to the polysilicon gate and R_s due to junction/contact resistance. In the forward active (saturation) region, the input capacitance is given by $C_{gs} = \frac{2}{3}C_{ox} \cdot W \cdot L$. The resistance r_o is due to channel length modulation and other short channel effects (such as DIBL). For a real transistor, layout parasitics increase the capacitance values but they are often lumped into the intrinsic model.

You're probably more familiar with the simplified FET model without gate and source resistance, as shown in Fig. 8.3a. Although these resistances are small, they play a crucial role when we analyze the power gain of the device. If we ground the bulk node, such as a common source amplifier, we can eliminate a lot of clutter since the g_{mb} generator is shorted (Fig. 8.3b).

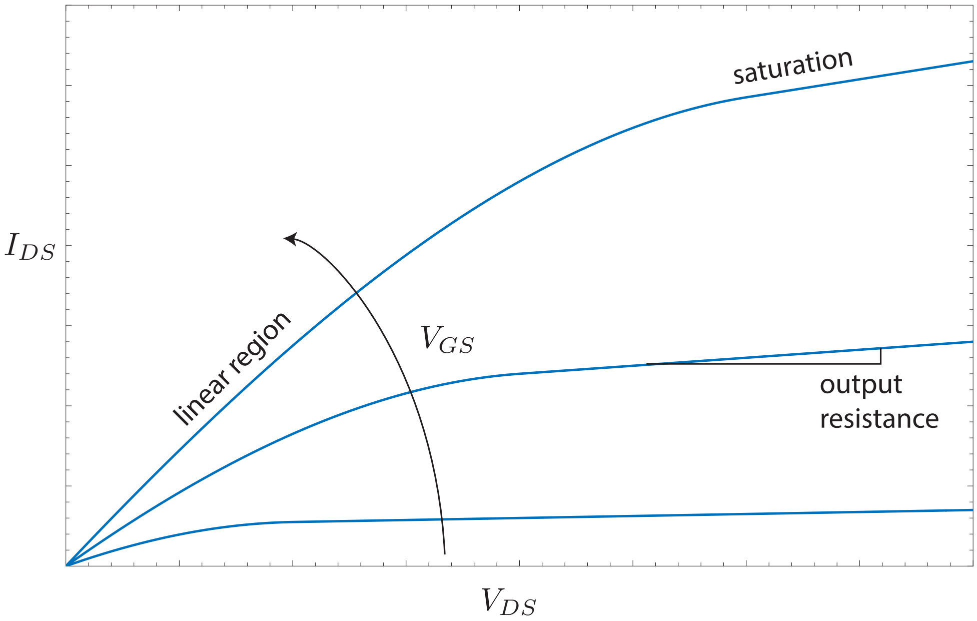


Figure 8.4: The I_{DS}-V_{DS} curve of a square-law MOSFET.

8.2.1 MOS Device Characteristics

The I_{DS}-V_{DS} curve of a typical square-law MOSFET is shown in Fig. 8.4. For a long channel FET, before "pinch-off", the device drain current responds nearly linearly with V_{DS}, hence the term "linear region".

$$I_{DS} = \mu C_{ox}\frac{W}{L}\left((V_{GS}-V_T)V_{DS}-\frac{V_{DS}^2}{2}\right) \tag{8.1}$$

Beyond pinch-off, when $V_{DS} = V_{GS} - V_T$, the current *saturates* and remains essentially constant. This is the saturation region.

$$I_{DS} = \frac{1}{2}\mu C_{ox}\frac{W}{L}\left((V_{GS}-V_T)^2(1+\lambda V_{DS})\right) \tag{8.2}$$

The variation of current in saturation is due to the output impedance of the device. Short channel devices have much stronger current variation. We also find that the current variation with V_{GS} is weaker than quadratic and the drain current is therefore lower than predicted by the long channel equations. This is partly due to the threshold voltage variation and reduced mobility.

The transconductance in saturation is given by

$$g_m = \frac{dI_{DS}}{dV_{GS}} = \mu C_{ox}\frac{W}{L}(V_{GS}-V_T)(1+\lambda V_{DS}) \tag{8.3}$$

or

$$g_m = \frac{2I_{DS}}{V_{GS}-V_T} = \frac{2I_{DS}}{\sqrt{\frac{2I_{DS}}{\mu C_{ox}\frac{W}{L}}}} \tag{8.4}$$

or

$$g_m = \sqrt{2\mu C_{ox}\frac{W}{L}I_{DS}} \propto \sqrt{I_{DS}} \tag{8.5}$$

The behavior of the device in the sub-threshold regime is also of some interest for analog and low performance RF applications. The device operates more as a lateral bipolar device where the *npn* junctions are formed by the source/substrate/drain combination. The body terminal is controlled through the gate through the field effect, rather than through an ohmic contact. In this regime the transconductance per current of the device is relatively large due to the "bipolar" exponential action. Unfortunately the speed of the device is relatively modest as the device f_T (see next section) is small.

8.2.2 MOS Unity Gain Frequency

The short circuit current gain of a device is given by

$$G_i = \frac{i_o}{i_i} \approx \frac{g_m}{j\omega(C_{gs} + C_{gd})} \tag{8.6}$$

The frequency of unity gain ω_T is given by solving $|G_i| = 1$

$$\omega_T = \frac{g_m}{C_{gs} + C_{gd}} \tag{8.7}$$

This frequency plays an important role in the frequency response of high speed amplifiers. Often there is a gain-bandwidth tradeoff related to ω_T

$$G \times BW = \omega_T \tag{8.8}$$

For a long-channel MOSFET we have the following relationship

$$\omega_T = \frac{g_m}{C_{gs} + C_{gd}} \approx \frac{\mu C_{ox} \frac{W}{L}(V_{GS} - V_T)}{\frac{2}{3} W \cdot L C_{ox}} \tag{8.9}$$

Canceling common factors we have

$$\omega_T = \frac{3}{2} \frac{\mu}{L^2}(V_{GS} - V_T) \tag{8.10}$$

We see that ω_T is bias dependent. The strong L^2 length dependence only holds for long-channel devices. Short channel devices, in the limit of velocity saturated operation, reduce to $1/L$ dependence. While mobility improves with V_{GS}, there is a point of diminishing returns since the mobility is a function of the field. At low fields mobility improves with the vertical field as the inversion layer "shields" carriers from Coulomb scattering sites. At high fields, though, the mobility drops due to increased surface scattering.

8.3 Bipolar Technology

In a bipolar junction (BJT) transistor, most of the transistor "action" occurs in the small vertical npn sandwich under the emitter, as shown in Fig. 8.5. The base width should be made as small as possible in order to minimize recombination. The emitter doping should be much larger than the base doping to maximize electron injection into the base. A SiGe HBT transistor behaves very similarly to a normal BJT, but has lower base resistance r_b since the doping in the base can be increased without compromising performance of the structure.

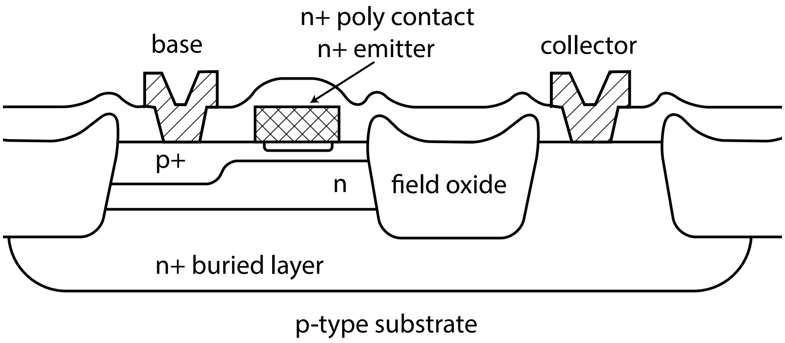


Figure 8.5: The cross-section of a bipolar junction transistor.

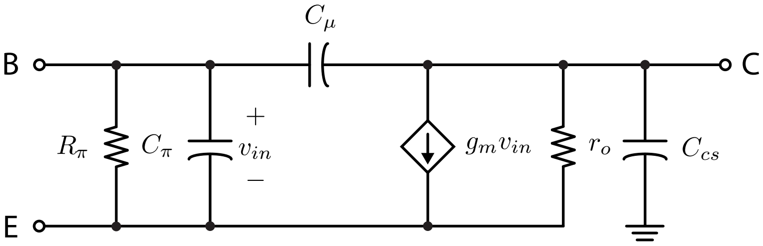


Figure 8.6: The small-signal equivalent circuit model for a BJT.

8.3.1 Bipolar Transistor Model

The core BJT model is similar to a FET small-signal model as shown in Fig. 8.6. The resistor R_π, though, dominates the input impedance at low frequency whereas at high frequencies, C_π dominates. C_μ arises from the collector-base reverse biased diode capacitance. C_{cs} is the collector to substrate parasitic capacitance. In some processes, this is reduced with an oxide layer. C_π has two components, due to the junction capacitance (forward-biased) and a diffusion capacitance

$$C_\pi = C_{bej} + C_{diff} \tag{8.11}$$

Due to Boltzmann statistics, the collector current is described very accurately with an exponential relationship

$$I_C \approx I_S e^{qV_{be}} kT \tag{8.12}$$

The device transconductance is therefore proportional to current

$$g_m = \frac{dI_C}{dV_{be}} = I_S \frac{q}{kT} e^{qV_{be}} kT = \frac{qI_C}{kT} \tag{8.13}$$

where $kT/q = 26\,\text{mV}$ at room temperature. Compare this to the equation for the FET. Since we usually have $kT/q < (V_{gs} - V_T)$, the bipolar has a much larger transconductance for the same current. This is the biggest advantage of a bipolar over a FET.

The generic figure shown in Fig. 8.7 represents both a FET and a bipolar at high frequency. Notice that this model holds when $R_\pi \gg \frac{1}{\omega C_\pi} = X_\pi$. Since

$$\frac{R_\pi}{X_\pi} = \omega R_\pi C_\pi = \omega \frac{\beta_0}{g_m} C_\pi \approx \beta_0 \frac{\omega}{\omega_T} \tag{8.14}$$

Say $\beta_0 = 100$ and the operating frequency is $\omega/\omega_T = 1/10$. Then we have $R_\pi/X_\pi = 100/10 = 10$.

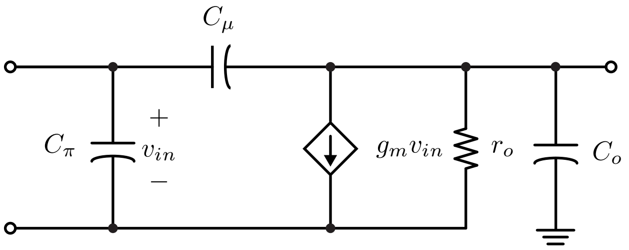


Figure 8.7: A generic small-signal transistor model.

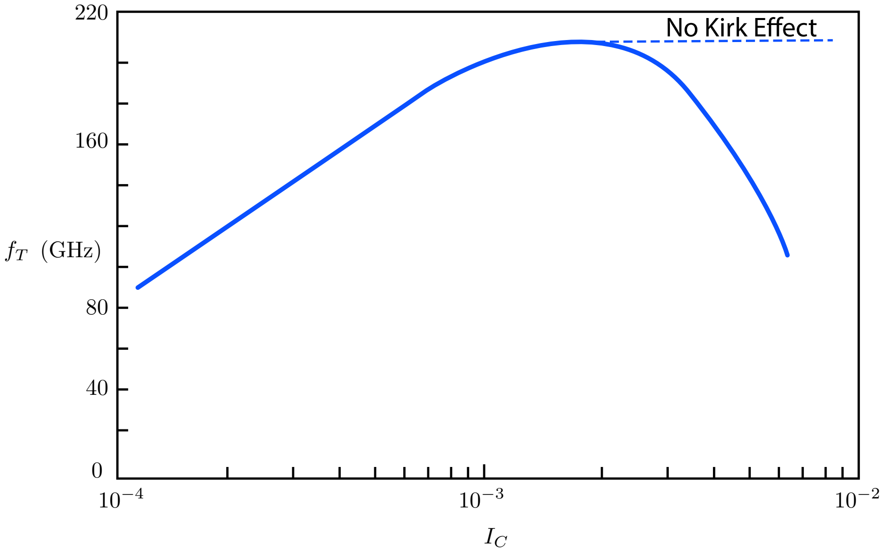


Figure 8.8: The unity gain frequency ω_T as a function of collector current.

8.3.2 Bipolar Unity Gain Frequency

Similar to a FET we have the following relationship

$$\omega_T = \frac{g_m}{c_\pi + C_\mu} \tag{8.15}$$

Expanding the denominator term

$$\omega_T = \frac{g_m}{C_{bej} + C_d + C_\mu} \approx \frac{g_m}{2C_{je0} + g_m \tau_F + C_{jc}} \tag{8.16}$$

The collector junction capacitance is a function of V_{bc}, or the reverse bias. To maximize ω_T, we should maximize the collector voltage. Re-writing the above equation

$$\omega_T = \frac{1}{\tau_F + \frac{2C_{je0}+C_{jc}}{g_m}} \tag{8.17}$$

We can clearly see that if we continue to increase I_C, then $g_m \propto I_C$ increases and the limiting value of ω_T is given by the forward transit time τ_F

$$\omega_T \to \frac{1}{\tau_F} \tag{8.18}$$

In practice, though, we find that there is an optimum collector current. Beyond this current the ω_T drops due to the Kirk Effect. It's related to the "base widening" due to high level injection.

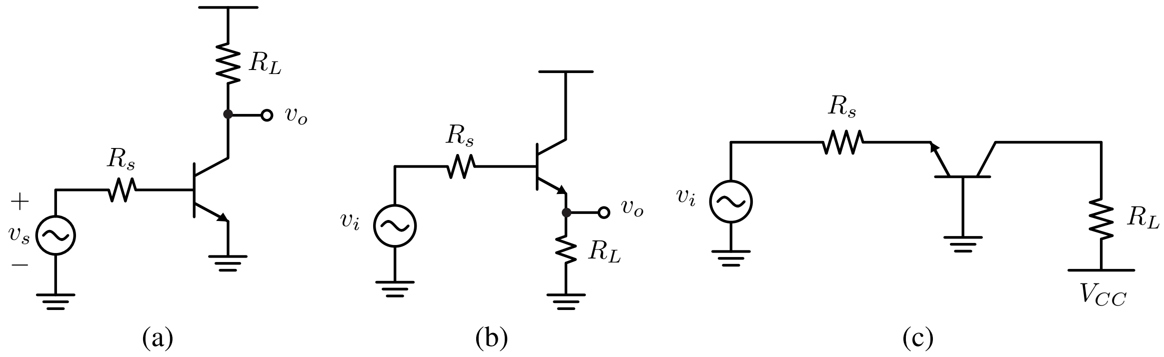


Figure 8.9: (a) Common-emitter, (b) common-collector (follower), and (c) common-base amplifiers.

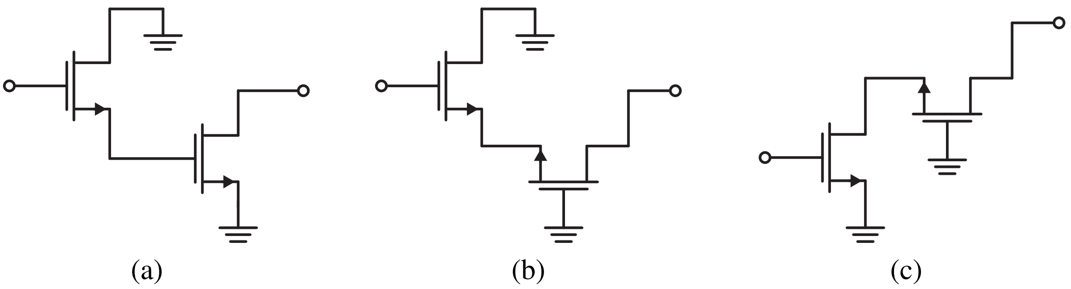


Figure 8.10: Wideband two-stage amplifiers.

8.4 Broadband Amplifier Blocks

Depending on the choice of common terminal, we can realize three well known amplifiers topologies shown in Fig. 8.9. The CE/CS amplifier has a small bandwidth due to the Miller feedback. In order to increase the bandwidth, we must keep the source resistance low. The CC/CD (or follower) is very wideband but only provides current gain. The CB/CG amplifier is also wideband (no Miller), but only offers voltage gain. It also features a small input impedance, which loads the driver. But in wideband applications this is a big advantage if the device is biased for impedance matching. The CE/CS offers the best power gain and noise figure, but bandwidth limitations are an issue.

To realize wider bandwidths without sacrificing gain, several two-stage amplifiers can be employed, as shown in Fig. 8.10. A source follower driving a common-source amplifier buffers the high source impedance and drives the common source amplifier with a low source impedance. A source follower driving a common gate amplifier boosts the input impedance. This is essentially a differential pair driven single-endedly. A common-source amplifier drives a common-gate amplifier, or a cascode amplifier. Miller effect is minimized by lowering the gain of the common-source stage.

These amplifiers are *broadband* amplifies because they realize gain over a wide bandwidth, typically from DC or low frequency up to the some fraction of the device f_T. A simple example will be used to demonstrate this fact.

The simple circuit shown in Fig. 8.11a is identified as a current mirror. This $1 \times N$ current mirror has broadband frequency response which can be illustrated with the equivalent circuit shown in Fig. 8.11b. The diode-connected device can be replaced with a conductance of value g_{m1} in shunt with the amplifier input capacitance C_{in}. If the current amplifier drives a low impedance load, the transfer function is given by

$$G_i = \frac{i_o}{i_s} = \frac{g_{m2}}{Y_{in}(s)} = \frac{g_{m2}}{g_{m1} + sC_{in} + N \times sC_{in}} \tag{8.19}$$

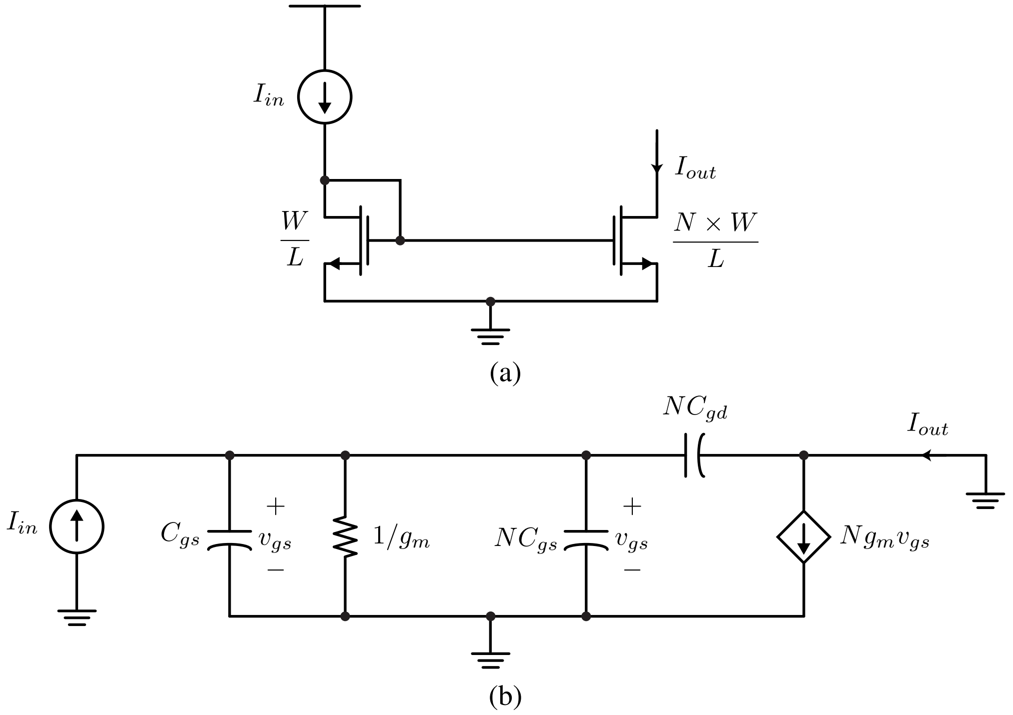


Figure 8.11: (a) A current mirror amplifier. (b) Small-signal equivalent of current mirror amplifier.

$$G_i = \frac{\frac{g_{m2}}{g_{m1}}}{1+(N+1)\frac{s}{\omega_T}} \tag{8.20}$$

Note that the transconductance of output device is N times larger since it can be thought of N devices in parallel. The complete transfer function is therefore

$$G_i = \frac{N}{1+\frac{s}{\omega_T/(N+1)}} \tag{8.21}$$

Which shows the bandwidth is a fraction $1/(N+1)$ of the gain, which implies the circuit has a fixed gain-bandwidth product

$$G_i \times \omega_{-3\text{db}} = \frac{N}{N+1}\omega_T \approx \omega_T \tag{8.22}$$

It's important to note that the above analysis holds only if we assume the load impedance is extremely low, ideally a short. If we connect a physical resistor to the output, the Miller effect will produce a significant feedback current which invalidates our assumptions. It is also interesting to note that the amplifier is large-signal linear.

The common-emitter (common source) amplifier shown in Fig. 9.10 has much smaller bandwidth due to Miller multiplication. The input capacitance is usually the dominant pole

$$\omega_0^{-1} \approx R_s(C_{in} + |A_v|C_\mu) \tag{8.23}$$

$$\omega_0^{-1} = R_s C_{in}(1+\mu|A_v|) \approx R_s C_{in}\mu|A_v| \tag{8.24}$$

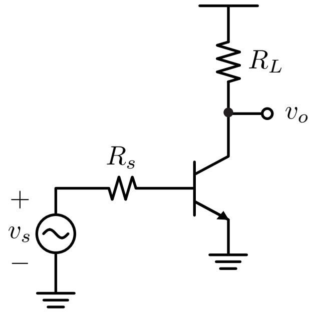


Figure 8.12: A common-emitter amplifier.

Assuming the voltage gain is given by the low-frequency value of $g_m R_L$, we have

$$\omega_0^{-1} = R_s C_{in} \mu g_m R_L = (g_m R_s)(g_m R_L) \frac{C_{in}}{g_m} \mu \tag{8.25}$$

$$\omega_0^{-1} = |A_v|^2 \frac{R_s}{R_L} \omega_T^{-1} \mu \tag{8.26}$$

The amplifier has a bandwidth reduction factor of A_v^2

$$\omega_0 \times |A_v|^2 = \omega_T \times \left(\frac{R_L}{R_s} \right) \times \frac{1}{\mu} \tag{8.27}$$

Let's work through a simple example. Say we need a gain of 60 dB ($A_v = 1000$) and $\frac{R_L}{R_s} = 2$. Assume that the technology has a capacitance ratio of $\mu = 0.2$:

$$\omega_0 |A_v|^2 = 10^6 \omega_0 = \omega_T \times 2 \times 5 \tag{8.28}$$

$$\omega_0 = \frac{\omega_T}{10^5} \tag{8.29}$$

Compare this to a current mirror amplifier. When we follow the "normal" gain-bandwidth tradeoff, we have

$$\omega_0 = \frac{\omega_T}{A_i} = \frac{\omega_T}{1000} \tag{8.30}$$

The common-base (common-gate) amplifier (Fig. 8.13) by contrast, is very wideband. Write KCL at base node of circuit (Fig. 8.14)

$$\frac{v_s + v_{in}}{R_s} + g_m v_{in} + s C_{in} v_{in} = 0 \tag{8.31}$$

$$v_s = -v_{in}(1 + g_m R_s + s C_{in} R_s) \tag{8.32}$$

And write KCL at the output node

$$(s C_o + \frac{1}{R_L}) v_o + g_m v_{in} + s C_\mu v_o = 0 \tag{8.33}$$

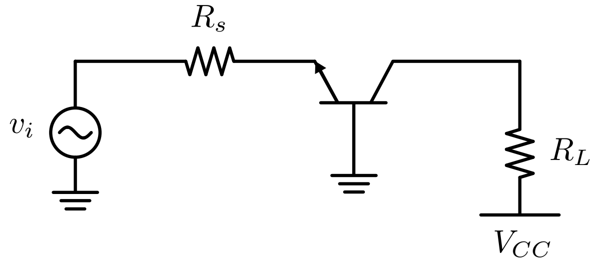


Figure 8.13: A common-base amplifier.

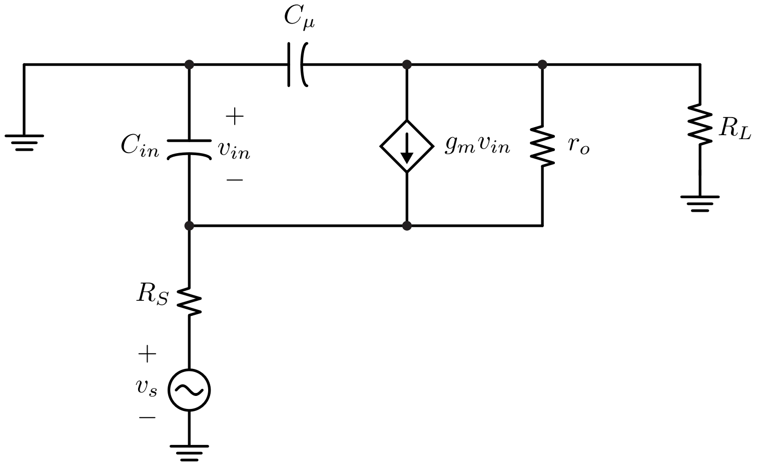


Figure 8.14: Small-signal equivalent circuit for common-base (common-gate) amplifier.

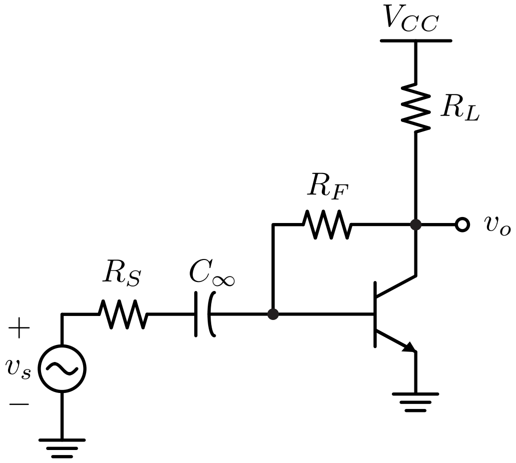


Figure 8.15: A local shunt-feedback amplifier.

$$v_o \left(\frac{1}{R_L} + s(C_o + C_\mu) \right) = -g_m v_{in} \tag{8.34}$$

The voltage gain is a product of two terms

$$A_v = \frac{v_o}{v_s} = \frac{-g_m R_L}{1 + s(C_o + C_\mu)R_L} \frac{v_x}{v_s} \tag{8.35}$$

$$A_v = \frac{G_m R_L}{(1 + s(C_o + C_\mu)R_L)(1 + sR_s \frac{C_{in}}{1+g_m R_s})} \tag{8.36}$$

Note the transconductance is degenerated, $G_m = g_m/(1 + g_m R_s)$. The input capacitance is also degenerated by the action of series feedback. Unlike a CE/CS ampilifier, the poles do not interact (due to absence of feedback capacitor). First let's take the limit of high loop gain, $g_m R_s \gg 1$

$$A_v = \frac{\frac{R_L}{R_s}}{(1 + s/\omega_T)(1 + s/\omega_L)} \tag{8.37}$$

where $\omega_L = \left((C_o + C_\mu)R_L\right)^{-1}$ is the pole at the output.

The common-base amplifier has the nice property that the input impedance is low (roughly $1/g_m$) and broadband, thus easily providing a termination to the driver (a filter, the antenna, or a previous stage). If we assume that $R_s = 1/g_m$, we have

$$A_v = \frac{\frac{1}{2} g_m R_L}{(1 + s/2\omega_T)(1 + s/\omega_L)} \tag{8.38}$$

The 3dB bandwidth is thus most likely set by the time constant at the load.

The shunt-feedback amplifier is a nice broadband amplifier building block. The action of the shunt feedback is used to lower the input impedance and to set the gain. The in-band voltage gain and input impedance is given by

$$A_v = \frac{-R_F}{R_s} \tag{8.39}$$

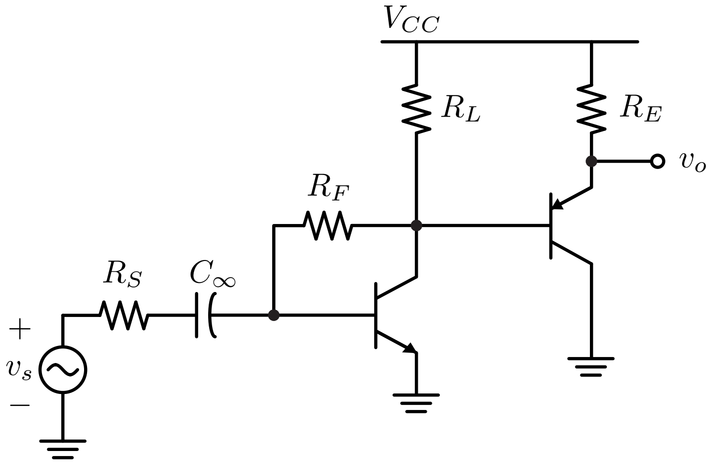


Figure 8.16: A common-collector amplifier drives a shunt-feedback amplifier to buffer the loading effects.

$$R_{in} = (1+\frac{R_F}{R_L})\frac{1}{g_m} \tag{8.40}$$

For an input match, $R_s = (1+\frac{R_F}{R_L})\frac{1}{g_m}$, or $g_m R_s = (1+\frac{R_F}{R_L})$. Since the voltage gain sets R_F, the input impedance match determines the required transconductance g_m (and hence the power dissipation). Generally a bipolar version will dissipate much less power due to the higher intrinsic g_m.

The amplifier is broadband and approximately obeys the classic gain-bandwidth tradeoff $A_v\omega_0 \approx \omega_T$. A zero-value time constant analysis identifies the dominant pole

$$\tau_1 = C_{in}\left(R_s||r_\pi||\frac{R_F(1+\frac{R_L}{R_F})}{1+g_m R_L}\right) \tag{8.41}$$

$$\tau_2 = C_\mu\left(R_F||\frac{R_L(R_s||r_\pi)}{R_s||r_\pi||\frac{1}{g_m}||R_L}\right) \tag{8.42}$$

$$\omega_{-3\text{dB}} \approx \frac{1}{\tau_1+\tau_2} \tag{8.43}$$

If the shunt-FB amplifier needs to drive a low impedance load, a broadband voltage buffer is needed. As shown in Fig. 8.16, an emitter follower (or source follower) provides the solution (note this is a fast pnp). Note the buffer is broadband (gain $\approx$ 1) and only loads the core amplifier by the degenerated input capacitance $C_{in2}/(1+g_{m2}R_E)$.

8.5 Narrowband Amplifier Blocks

The *RLC* loaded amplifier shown in Fig. 8.17 is called a tuned amplifier. A transconductance device drives a shunt *RLC* load which results in a voltage gain

$$A_v = -g_m Z(j\omega) = \frac{-g_m}{Y(j\omega)} \tag{8.44}$$

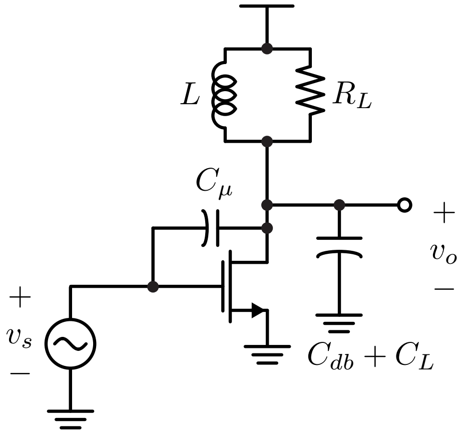


Figure 8.17: A tuned *RLC* amplifier.

The peak gain occurs at resonance

$$A_{v,\max} = -g_m R_{\text{eff}} \tag{8.45}$$

R_{eff} is the *loaded* resistance of the tank

$$R_{\text{eff}} = R_L || r_o || R_{x,L} || R_{x,C} \tag{8.46}$$

The tuned amplifier has a 3-dB bandwidth of ω_0/Q. If the Q is large, the amplifier is narrowband. In essence, we can view this amplifier has an active filter. In order to maximize the gain, we employ high-Q inductors and capacitors in the load and omit the explicit load resistance R_L. The peak gain is thus

$$A_{v,\max} \approx -g_m (R_{x,L} || R_{x,C}) \tag{8.47}$$

Assuming the Q factor is dominated by the inductor, a good assumption for monolithic IC inductors, we have

$$A_{v,\max} \approx -g_m R_{x,L} = -g_m Q_L \omega L \tag{8.48}$$

8.5.1 Tuned Load Calculation

We use the series to parallel transformation to calculate the effective shunt resistance due to the inductor (see Sect. 6.1.1)

$$R_{\text{eff}} = (1 + Q_L^2) R_{x,L} \approx Q_L^2 R_{x,L} = Q_L^2 \frac{\omega L}{Q_L} = Q_L \times \omega L \tag{8.49}$$

The gain is maximized at a fixed bias current and frequency by maximizing the product $Q_L \times L$. So in theory, there is no limit to the voltage gain of the amplifier as long as we can maximize the quality factor Q_L.

Note that the capacitance of the circuit is not detrimental since it is resonated away with the shunt inductance. In other words L is chosen such that

$$L = \frac{1}{\omega_0^2 C_{\text{eff}}} \tag{8.50}$$

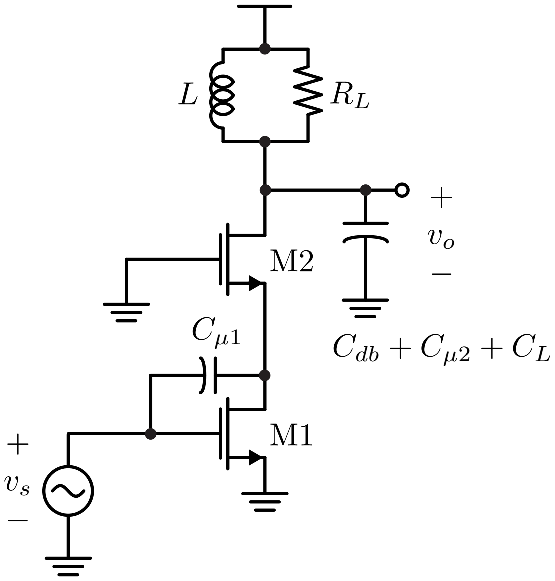


Figure 8.18: A tuned cascode amplifier.

where $C_{\text{eff}} = C_{db} + (1 - |A_v^{-1}|)C_\mu + C_L$. The ability to tune out the parasitic capacitances in the circuit is a major advantage of the tuned amplifier. This is especially important as it allows low-power operation. Another important advantage of the circuit is that there is practically no DC voltage drop across the inductor, allowing very low supply voltage operation.

Another less obvious advantage is the improved voltage swing at the output of the amplifier. Usually the voltage swing is limited by the supply voltage and the $V_{ds,sat}$ of the amplifier. In this case, though, the voltage can swing above the supply. Since the average DC voltage across an inductor is zero, the output voltage can swing around the DC operating point of V_{dd}. This is a major efficiency boost for the amplifier and is an indispensable tool in designing power amplifiers and buffers.

8.5.2 Cascode Tuned Amplifier

A cascode tuned amplifier shown in Fig. 8.18, has several advantages. In addition to boosting the output impedance, thus maximizing the Q of the load, the cascode device solves a major stability problem of the amplifier. We'll show that a feedback C_μ path can easily lead to unwanted oscillations in the amplifier. The cascode tuned amplifier effectively has zero feedback and thus is much more stable. The loss in voltage headroom is a small price to pay considering the improved headroom afforded by the inductor.

Bandwidth

It's interesting to note that the bandwidth of the tuned amplifier is still determined by the RC time constant at the load

$$BW = \frac{\omega_0}{Q} = \frac{\omega_0}{\omega_0 RC} = \frac{1}{RC} \tag{8.51}$$

The ultimate sacrifice for high frequency operation in a tuned amplifier is that the amplifier is narrow-band with zero gain at DC. In fact, the larger the Q of the tank, the higher the gain and the lower the bandwidth.

How high can we go? To win some of the bandwidth back requires other techniques, such as shunt peaking and distributed amplifiers. But can we tune out parasitics and design amplifiers

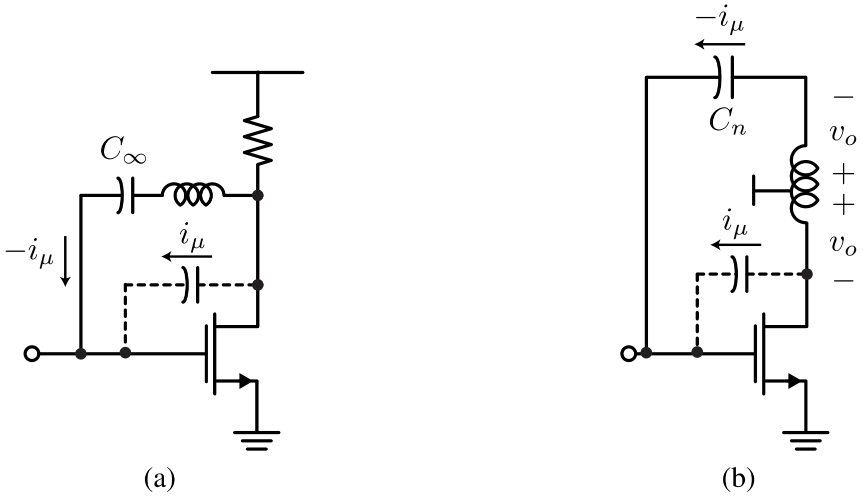


Figure 8.19: (a) A narrowband neutralization technique. (b) A tapped-inductor or transformer based neutralization approach.

operating at arbitrarily high frequency? Based on the simple analysis thus far, it seems that for any given frequency, no matter how high, we can simply absorb the parasitic capacitance of the amplifier with an appropriately small inductor (say a short section of transmission line) and thus realize an amplifier at an arbitrary frequency. This is of course ludicrous and we'll re-examine this question later.

8.5.3 Neutralization

For a non-unilateral transistor, we can try to come up with ways to eliminate the feedback through C_μ in order to improve the gain and mitigate the de-stabilizing effect of feedback at high frequencies. To do this, ideally we'd like a negative capacitor but this is not physical. An inductor can be used as shown to resonate the capacitance at a single frequency (Fig. 8.19a). Another technique uses a tapped inductor as shown in Fig. 8.19b. Since

$$i_\mu = (v_o - v_{in})sC_\mu \approx v_o sC_\mu \tag{8.52}$$

The voltage at the output of the other tapped inductor is phase inverted, so the current in the external C_μ is given by

$$i'_\mu = (-v_o - v_{in})sC_\mu \approx -v_o sC_\mu \tag{8.53}$$

thus "neutralizing" the feedback current

An elegant strategy for neutralization is a to employ the natural signal inversion in a differential pair as shown in Fig. 8.20. Not only is the output phase inverted, but the input is also inverted so the neutralization occurs independent of the gain (unlike the previous strategies).

8.6 Amplifier Design Techniques

The design of amplifiers begins with the specifications for the performance, including

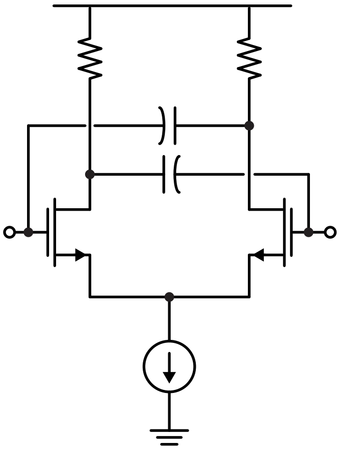


Figure 8.20: The neutralization of a differential pair amplifier through cross-coupling.

- Amplifier power gain. Most RF amplifiers are specified in terms of power gain and the load/source impedance are known (e.g. $Z_0 = 50\,\Omega$). In an integrated amplifier, the internal impedance may be higher and unmatched, in which case voltage gain might be more appropriate.
- Amplifier bandwidth. We usually desire operation over a specified bandwidth, where we expect the gain and group delay to be relatively flat (to avoid distortion) over a certain range. Two common amplifiers are baseband amplifiers, which must operate from DC or low frequency up to a given bandwidth (say DC - 100 MHz), and narrowband amplifiers which operate at RF (say 2.4 GHz). Narrowband amplifiers operate at a given center frequency and realize selectivity through the use of resonant circuitry.
- Input and output match are usually specified as the maximum tolerable reflection coefficient at the input and output (e.g. $S_{11} < -10\,\text{dB}$). Matching is important in order to extract the maximum power from a source (antenna), to properly terminate a transmission line (otherwise the power gain will depend on the length of the line which changes with frequency), or to provide the proper termination for a filter for proper filter response.
- Amplifier stability. A robust amplifier should be stable over all frequency ranges and over process and temperature. Process and temperature variations cause the operating point to shift and thus stability should be checked under these conditions. Absolute stability ($K > 1$) implies that the amplifier is stable for any source or load impedance. A conditionally stable amplifier ($K < 1$) will become unstable if the load/source take on particular values. The stability circle plot on the Smith Chart shows the regions of instability. If the load and source are fixed (say at Z_0), then a conditionally stable amplifier may be acceptable. The designer should ensure that the unstable region is far from the origin of the Smith Chart and does not come too close under all conditions (frequency/temperature/process/bias voltage variations). Stability versus load variations is often specified through SWR, or the standing wave ratio. If an amplifier is stable over an $N : 1$ SWR, that means the magnitude of the load can vary by a factor of N above or below the nominal value.

Other important specifications include:

- Amplifier noise figure for receiver applications. Noise figure is a measure of how much noise the amplifier adds to the signal, which degrades the signal-to-noise ratio (SNR) of the signal, resulting in lower receiver sensitivity. Noise is most important when dealing with weak or small signals, when the noise signal amplitude is a substantial fraction of the input signal. We will cover this topic in depth in Chapter **??**.

- Amplifier distortion generated by active device non-linearity. Distortion specifications include harmonic distortion (HD), which occur at harmonics of the input frequency, and intermodulation (IM) distortion, which also occur in-band, or near the operating frequency. In RF applications, IM distortion is much more important since harmonic distortion can be filtered out. These distortion products must be kept sufficiently small so that the amplified signal is not severely distorted. The strength of the distortion signals increases rapidly with the signal amplitude (faster than linear), which means that we mostly care about distortion when the signal is strong. We'll cover distortion in depth in Chapter **??**.
- Amplifier performance under process and temperature variations. In practice every amplifier will perform slightly differently due to inevitable variations in component parameters, especially active devices. Active devices are especially sensitive to temperature variations, and much effort is typically dedicated in designing a biasing network to cope and compensate for such variations.
- Amplifier efficiency. The efficiency is determined by comparing how much power is delivered to the load divided by the DC power consumption of the amplifier. This metric is most important for power amplifiers which consume significant power and it's desirable to use as much of this power as possible (power transmitted through the antenna versus wasted power converted to heat). If the gain is low, the input power should also be counted in the *power added efficiency*

$$\eta = \frac{P_L - P_{in}}{P_{DC}} \tag{8.54}$$

  We'll return to the issue of efficiency in Chapter **??**.

The next step in the design is the selection of the active devices (technology) and the bias point for the transistors. The bias point must be chosen carefully to meet power dissipation constraints, imposed either by physical limitations, such as thermal and DC voltage headroom, or to minimize the power consumption. Once a device has been selected, the designer should plot the maximum achievable gain (G_{max}/G_{MSG}) for the device over the frequency range of interest. If the device is unconditionally stable, then in practice we can come close to this maximum gain, but keep in mind that there is some loss in the matching networks at the input and output of the amplifier. If the device is conditionally stable, the designer may intentionally introduce loss or feedback in the device to stabilize the device. Other techniques such as neutralization through feedback or the use of a compound device (such as a cascode) are other options. If the required gain cannot be realized with a single device, then two or more stages are required. But keep in mind that the bandwidth may be limited by the matching networks. If the device optimal source/load impedances are vastly different than Z_0, then the Q of the required matching networks will be very large, which will limit the bandwidth.

An important consideration in the design of amplifiers is the use of feedback. If the loop gain is sufficiently high, then negative feedback amplifier performance is determined by the passive feedback network rather than by the active component parameters, leading to less variation with process and temperature. Feedback amplifiers may introduce instability at high frequency due to the phase shift through the device, which converts negative feedback into positive feedback, de-stabilizing the circuit. At high frequencies it's common to avoid this effect by only using "local" feedback, or feedback around a single device (shunt or series feedback). In addition, in narrowband applications it's advantageous to use reactive components (for instance series feedback using an inductor) so as to introduce less noise into the circuit. Most importantly, even if the designer does not wish to employ feedback, at high frequency it's hard to avoid parasitic feedback, which occurs due to small feedback capacitors (such as the Miller capacitor) in active devices and due to board and package parasitics. The package introduces inductive parasitics in the form of lead inductance around the transistor and mutual inductive coupling between the leads of the transistor. If the

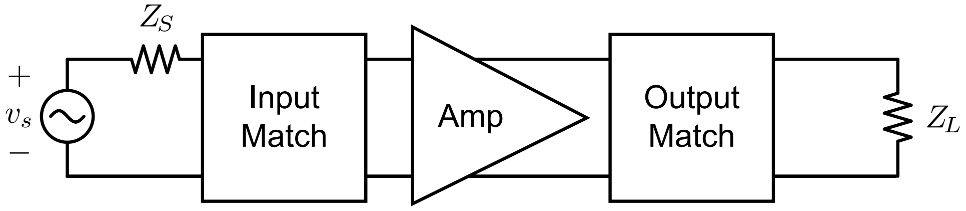


Figure 8.21: The design of an RF amplifier can be factored into the design of the input matching network, the bias of the amplifier, and the output match network.

amplifier is fully integrated and employs sufficient bypass capacitors, then package parasitics are only important when making connections to the external world, or at the input and output of the amplifier.

Once the amplifier transistors have been biased and stabilized, one selects the source/load impedance to achieve the desired performance. Typically driving the amplifier directly from a source and load of Z_0 impedance will result in a gain which is too low. For instance, the voltage gain of a simple resistively loaded amplifier contains a term $g_m Z_L$. If the load impedance $Z_L = Z_0$, then the voltage gain will be low since typically $Z_0 \sim 50\,\Omega$. On the other hand, if a matching network is used to raise the load impedance to r_o (matched) or a large value (voltage amplifier), then much higher gain can be realized. The source and load are therefore transformed into the desired impedances through matching networks. Matching networks play a crucial role in the design of RF amplifiers, since once the device bias is chosen, the only other degree of freedom is the matching network. In Fig. 8.21 we show an input matching network cascaded with a two-port amplifier followed by an output matching network. The overall gain of the two-port can be written as

$$G_T = \frac{1-|\Gamma_S|^2}{|1-\Gamma_{in}\Gamma_S|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \tag{8.55}$$

which can be viewed as a product of the action of the input match “gain", the intrinsic two-port gain $|S_{21}|^2$, and the output match “gain". Since the general two-port is not unilateral, the input match is a function of the load. Likewise, by symmetry we can also factor the expression to obtain

$$G_T = \frac{1-|\Gamma_S|^2}{|1-S_{11}\Gamma_S|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-\Gamma_{out}\Gamma_L|^2} \tag{8.56}$$

In the case of a unilateral amplifier, the above equations simplify to the product of three independent terms ($\Gamma_{in} = S_{11}$)

$$G_T = \frac{1-|\Gamma_S|^2}{|1-S_{11}\Gamma_S|^2} \times |S_{21}|^2 \times \frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} = M_1 \times |S_{21}|^2 \times M_2 \tag{8.57}$$

The above equation clearly shows the role of the input and output matching network. For a conjugate match, $\Gamma_S = S_{11}^*$, which means the maximum gain for the input matching network is given by

$$M_{1,\max} = \frac{1}{1-|S_{11}|^2} \tag{8.58}$$

If the amplifier has $S_{11} \lesssim 1$, we can improve the transducer gain considerably by matching the input. A similar consideration applies to the output of the amplifier. The design of non-unilateral amplifiers is more complicated but often we can ignore the reverse feedback if $S_{12} \simeq 0$. The

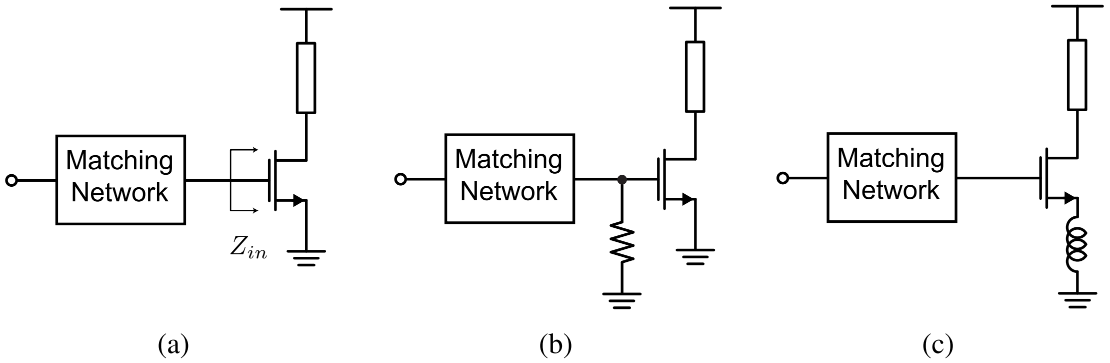


Figure 8.22: (a) Matching networks often must convert a predominantly imaginary load impedance to a real value. (b) A simple solution is to simply terminate the matching network with a physical resistor. (c) A more elegant solution uses a feedback synthesized resistor input match.

Unilateral Figure of Merit UFM is a good metric to test determine the error that would result if you make this assumption [**Pozar**]

$$UFM = \frac{|S_{11}||S_{22}||S_{12}||S_{21}|}{(1-|S_{11}|^2)(1-|S_{22}|^2)} \tag{8.59}$$

which sets an upper and lower bound for the error in gain under the unilateral assumption

$$\frac{1}{1+UFM^2} < \frac{G_T}{G_{TU}} < \frac{1}{1-UFM^2} \tag{8.60}$$

where G_{TU} is the calculated value of gain by neglecting S_{12}.

8.6.1 Reactive Series Feedback

In our discussion of matching networks, we considered networks that transform from a given source impedance to a given load impedance. Consider now the load shown in Fig. 8.22a, the input of an active device. At moderate frequencies the input impedance is dominated by C_{gs}. We need to somehow transform the input capacitance to a real load resistance. Any real MOS amplifier has a real component to the input impedance and thus there is a finite real component to the input impedance. If the transistor layout has ample fingers to minimize the physical polysilicon (or metal) gate resistance, the remaining gate induced channel resistance is given by $1/5g_m$. Thus the Q factor of the input of the MOS transistor is given by

$$Q_{gate} \approx \frac{5g_m}{\omega C_{gs}} = 5\frac{\omega_T}{\omega} \tag{8.61}$$

At moderate frequencies $\omega \ll \omega_T$, this is a high Q input impedance. If we resonate out this capacitance with a shunt inductor, the resulting shunt resistance is $Q^2 R_i$ is too large to match to the low source resistance. On the other hand, if use a series inductor the input resistance is simply R_i at resonance, which is too small to match. So what do we do?

We could explicitly add a resistor to the gate, as shown in Fig. 8.22b, but this will add extra noise to the circuit. A more elegant non-obvious solution is to add an inductor to the source of the amplifier, shown in Fig. 8.22c. To see the benefit, consider the general case of a degeneration impedance Z connected to the source. It's easy to show that the input impedance becomes (neglecting C_{gd})

$$Z_{in} = Z + \frac{1}{j\omega C_{gs}} + \frac{g_m Z}{j\omega C_{gs}} \tag{8.62}$$

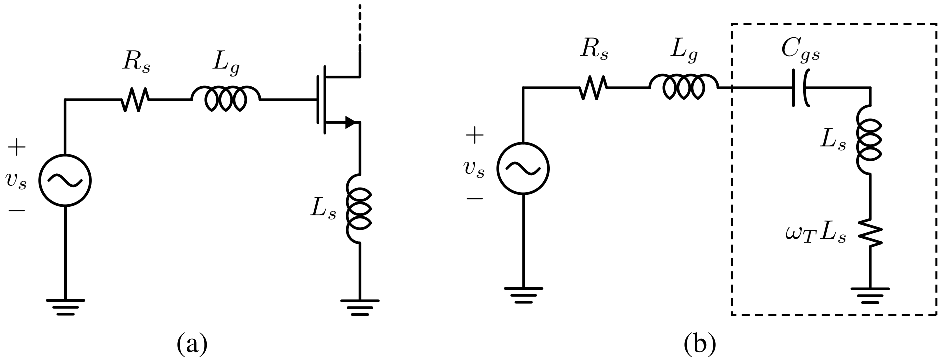


Figure 8.23: (a) The complete input matching network requires a gate inductor L_g to resonate with the input capacitance C_{gs}. (b) The equivalent circuit for the input match is a series *RLC* circuit.

The action of the feedback produces the third term, which can be programmed appropriately. Note that if $Z = j\omega L_s$, the third term becomes purely real and independent of frequency

$$\Re(Z_{in}) = R_{in} = \frac{g_m L_s}{C_{gs}} = \omega_T L_s \tag{8.63}$$

By simply controlling value of L_s, we can control the input impedance. We can also vary the ω_T of the device by placing extra capacitance in shunt with C_{gs}.

It's interesting to observe that the source impedance in effect drives a series *RLC* circuit, shown in Fig. 8.23. The voltage gain can be calculated by noting that the voltage v_{gs} is the voltage across a series resonant capacitor, which means that it's Q times as large as the voltage across the source resistor. For an input match, $v_{R_S} = \frac{1}{2}v_s$, so the voltage gain at resonance is given by

$$v_o = -g_m R_L v_{gs} = -g_m R_L Q \times \frac{v_s}{2} \tag{8.64}$$

or

$$A_v = -\tfrac{1}{2} g_m R_L Q \tag{8.65}$$

The bandwidth of matching stage of the inductively degenerated amplifier is set by Q factor of the input. Since the source impedance is fixed, there is little freedom in controlling the Q factor of the input stage. In most designs, the Q is fairly low and thus the input stage is relatively wideband.

Example 15:A Discrete Transistor Amplifier Design

Suppose we wish to design a discrete amplifier meeting following specifications: Power Gain $G_p > 12$ dB, S_{11} and $S_{22} < -10$ dB, operating at a center frequency of 1.2 GHz with a bandwidth of at least 200 MHz. The current consumption of the amplifier should not exceed 5 mA. One may elect to use a single stage or a two-stage amplifier. One can also optionally employ feedback for DC biasing, stability, and matching. All of these considerations depend strongly on the choice of technology. For a discrete transistor, one must carefully factor the device parasitics.

The SPICE model for the transistor is usually given by the manufacturer, for example for the NXP transistor shown in Fig. 8.24, is found at `http://www.nxp.com/models/`

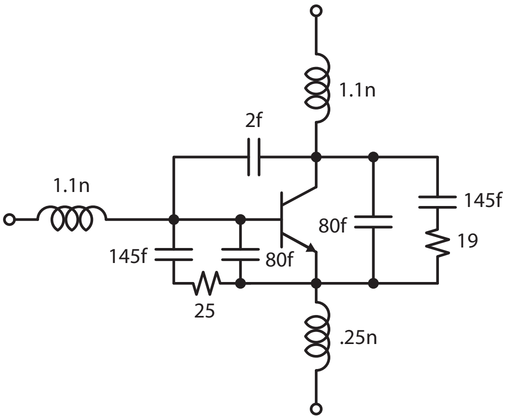


Figure 8.24: Equivalent circuit for packaged bipolar transistor.

`spicespar/data/BFG403W.html`. Notice that the transistor model is embedded in a package model which includes the parasitic effects of the package and bond wires, including lead inductance and capacitance, and mutual inductance and capacitance. The component values for the BFG40Wp transistor are given below.

```
subckt BFG403Wp base collector emitter inh_bulk_n
    Q1 (net17 net029 net034 inh_bulk_n) BFG403W region=fwd
    I10 (net4 net034) res r=19
    I9 (net6 net034) res r=25
    I8 (net17 net4) cap c=1.45e-13
    I7 (net029 net6) cap c=1.45e-13
    I6 (net17 net034) cap c=80e-15
    I5 (net029 net034) cap c=80e-15
    I4 (net029 net17) cap c=2e-15
    I3 (net034 emitter) ind l=0.25e-9
    I2 (collector net17) ind l=1.1e-9
    I1 (base net029) ind l=1.1e-9
ends BFG403Wp
```

```
.MODEL BFG403W NPN
+ IS = 5.554E-18
+ BF = 145
+ NF = 0.9934
+ VAF = 31.12
+ IKF = 35.75E-03
+ ISE = 3.535E-14
+ NE = 3
+ BR = 11.37
+ NR = 0.985
+ VAR = 1.874
+ IKR = 14.3E-03
+ ISC = 5.708E-17
+ NC = 1.546
+ RB = 122.38
+ IRB = 0
+ RBM = 52.45
+ RE = 1.511
+ RC = 15.119
+ CJE = 3.661E-14
+ VJE = 0.9
+ MJE = 0.3456
+ CJC = 1.621E-14
+ VJC = 0.5569
+ MJC = 0.2079
+ CJS = 7.859E-14
+ VJS = 0.4183
+ MJS = 0.2391
+ XCJC = 0.5
+ TR = 0.0
+ TF = 4.122E-12
+ XTF = 68.2
+ VTF = 2.004
+ ITF = 0.1796
+ PTF = 0
+ FC = 0.5501
+ EG = 1.11
+ XTI = 3
+ XTB = 1.5
```

It is very important to take these parasitics into account at high frequency in order to properly predict the performance of the practical amplifier. In addition to the package parasitics, you must be sure to include an estimate of the component parasitics, such as finite inductor Q. This will impact the matching network, gain, and stability of your amplifier. Finally, you must estimate the board level parasitics in the design. For instance, if you short the emitter of the amplifier, it actually must travel through the package and then through the board. For example, if a 0603 footprint is added for series feedback, this component introduces inductance even if it is shorted with a zero ohm resistor. The via also contributes inductance.

Here are some steps in the design.

1. Simulate the f_T of the transistor versus collector current. Find the current where f_T is maximized and simulate the f_{max} at this bias point. What is the highest frequency one can use this transistor and realize a power gain of 12 dB?
2. Plot the BJT transistor's maximum stable gain (MSG) and stability factor at the design bias point. Be sure to include the package parasitics. At what frequency is the transistor unconditionally stable?
3. Design your amplifier with equations as much as possible. You may find it convenient to extract a hybrid-π model at a single frequency from the device Y parameters. Include calculations and simulations results used to arrive at the design. Do not use "SPICE monkey" techniques in your design! Make sure the required component values are realizable in the footprint (for example 0603).
4. Design a bias network for your amplifier. Common approaches include base resistor dividers with emitter series feedback or self-biasing through a shunt feedback resistor R_f. Size your biasing resistors so that they do not interfere with the amplifier at high frequencies. Use bypass capacitors where appropriate. Check the stability of the biasing scheme by varying the BJT transistor parameters (β_0 and I_S) and verify that the amplifier bias remains relatively constant.
5. Simulate your amplifier (ADS or SpectreRF) and verify the performance. Include plots of the power gain, stability factor, input match, and output match. Identify the bandwidth of the amplifier. Be sure that you simulation includes the bandwidth of the input/output match (simulate port-to-port rather than S_{21}).
6. Compare the gain of the amplifier to $|S_{21}|^2$. How much "gain" do you realize with the input and output match?
7. How does the amplifier's overall gain compare to the MSG of the device? How much gain is lost in the input/output match (due to finite Q components)?
8. Simulate your amplifier using Monte-Carlo analysis by varying the components by 20%. Plot the amplifier stability under variations.

Building A Prototype

The board layout for the single-stage amplifier is shown in Fig. 8.25. When soldering active devices such as the BJT transistor, be sure to strap yourself to ground to avoid electrostatic discharge from damaging the device. Notice that the board has room for a lot more components than you may actually need to give you the maximum flexibility for your design.

1. Solder the components onto board and make sure the amplifier biases correctly. Solder the transistor chip last to minimize risk of damage to the integrated circuit. Check the device V_{BE} and V_{CE} to ensure proper biasing conditions.
2. Be sure to include bypass capacitors from DC points to ground. Solder the

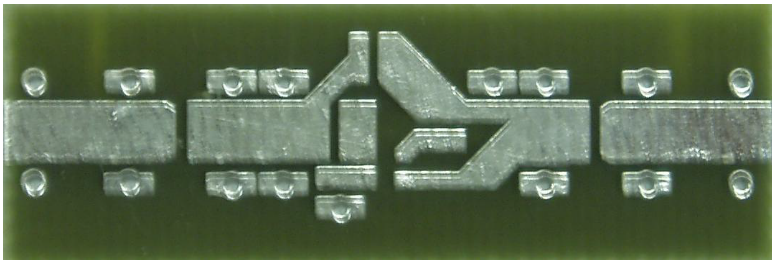


Figure 8.25: PCB for the amplifier includes footprints for input/output matching, feedback, and biasing.

capacitors on the back-side of the board at points to minimize the inductance of the DC point. Since large capacitors self-resonate at lower frequencies, below the desired operating frequency, use several parallel capacitors (1pF - 1μF) to realize a broadband low impedance at the supply.

3. Bias up the amplifier and ensure that the DC operating point matches expectations. Check the DC point at each node and calculate the voltage V_{BE} and V_{CE} to verify the amplifier is in the correct operating region. Do not proceed until the amplifier is operating correctly.
4. If needed, vary the DC bias to match your simulation results for current.
5. Measure and record the amplifier frequency response (S_{11}, S_{22}, S_{21} and S_{12}) on the VNA over the desired frequency range.
6. Measure and record the amplifier frequency response over a broad frequency range (DC to the maximum available frequency of the VNA). Plot the amplifier stability factor over this range.
7. Plot the input/output impedance (magnitude/phase) of the amplifier over the desired frequency range.
8. From measured data, what's the G_{max} of the amplifier at the center frequency?
9. Use matching feature of VNA to improve gain of the amplifier. Find the optimal source/load impedance and virtually "embed" (through simulation on the VNA) these impedances into your amplifier and plot the results.
10. Partner up with another group and measure the cascaded performance of your amplifiers. Measure the small-signal gain and bandwidth. Verify that the measured performance matches with the calculated cascade performance, particularly the overall gain and S_{21}. If the amplifiers are not well matched, be sure to include the effect of mismatch in your gain calculation.
11. Using the spectrum analyzer observe the output spectrum. Insert a weak tone at 900 MHz (-60 dBm) at the input and observe the output spectrum. Vary the input power and measure the power at the fundamental and a few harmonics. C

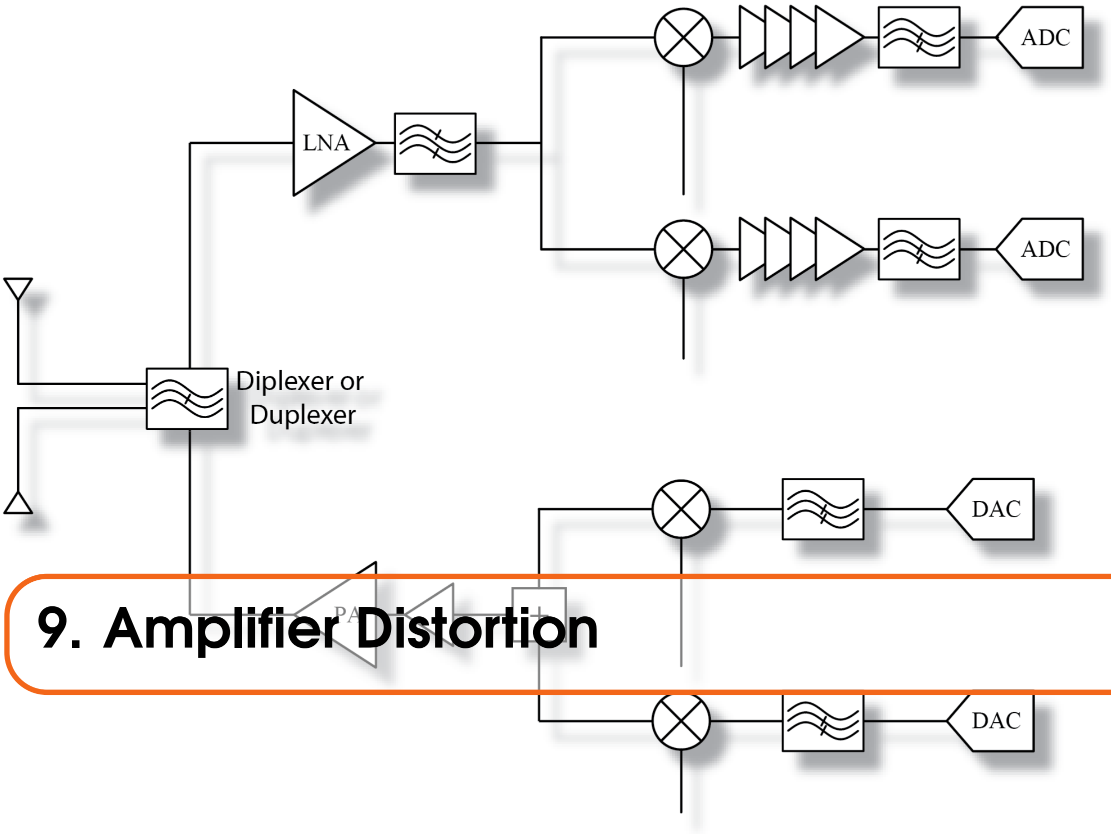
Save your amplifier board for future experiments!

Important Lessons

1. Explain the importance of bypass capacitors in the design. Why does the supply V_{CC} need to be bypassed to ground? Where is the best location to place bypass capacitors?
2. Compare measurement and simulation results by overlapping the measured S_{11}, S_{22} and S_{21} of your amplifier with the simulations. Explain differences.
3. Explain the significance of S_{12} in the design. Compare the measured and simu-

lated S_{12}, especially at higher frequencies. Explain the mismatch between the simulation and measurements.

4. Which board level or component level parasitics were the most detrimental? Explain.
5. Modify schematic to match measurements. You should use your knowledge of the board level parasitics and component parasitics. Make sure you understand the various measurements of the board level parasitics. The component models can also be obtained from the manufacturer.



9. Amplifier Distortion

9.1 Introduction to Distortion

Up to now we have treated amplifiers as small-signal linear circuits. Since transistors are non-linear, this assumption is only valid for extremely small signals. Consider a class of *memoryless* non-linear amplifiers. A memoryless system can be described by an instantaneous input/output relation (see Fig. 9.1. In other words, the output at a given time only depends on the input, and not any past history. In other words, let's neglect energy storage elements (inductors, capacitors). We also assume the input/output description is sufficiently smooth and continuous as to be accurately described by a power series

$$s_o = a_1 s_i + a_2 s_i^2 + a_3 s_i^3 + \dots \tag{9.1}$$

Bipolar Transistor Distortion

For instance, for a BJT (Si, SiGe, GaAs) operated in forward-active region, the collector current is a smooth function of the voltage V_{BE} as shown in Fig. 9.2a.

$$I_C = I_S e^{qV_{BE}/kT} \tag{9.2}$$

As illustrated in Fig. 9.2b, we can shift the origin by eliminating the DC signals, $i_o = I_C - I_Q$. The input signal is then applied around the DC level $V_{BE,Q}$. Note that an ideal amplifier has a perfectly linear line, and so the BJT amplifier is only linear for small signals. We shall quantify what we mean by "small" shortly.

MOSFET Distortion

In general for a MOSFET $I_{DS} = f(V_{GS}, V_{DS})$ (neglecting the body effect) but if the device is biased in the saturation regime, the output current is only a weak function of V_{DS}. The long-channel device follows the square law relation (neglecting bulk charge effects) (Fig. 9.3a)

$$I_D = \tfrac{1}{2}\mu C_{ox}\frac{W}{L}\left(V_{GS} - V_t\right)^2\left(1 + \lambda V_{DS}\right) \tag{9.3}$$

This is assuming the device does not leave the forward active (saturation) regime (Fig. 9.3b). Note that the device operation near threshold is not captured by our simple square-law equation. The I-V

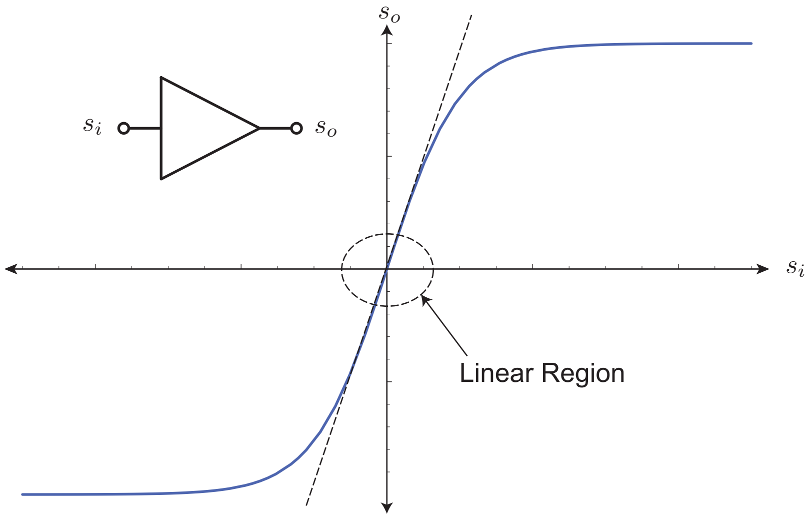
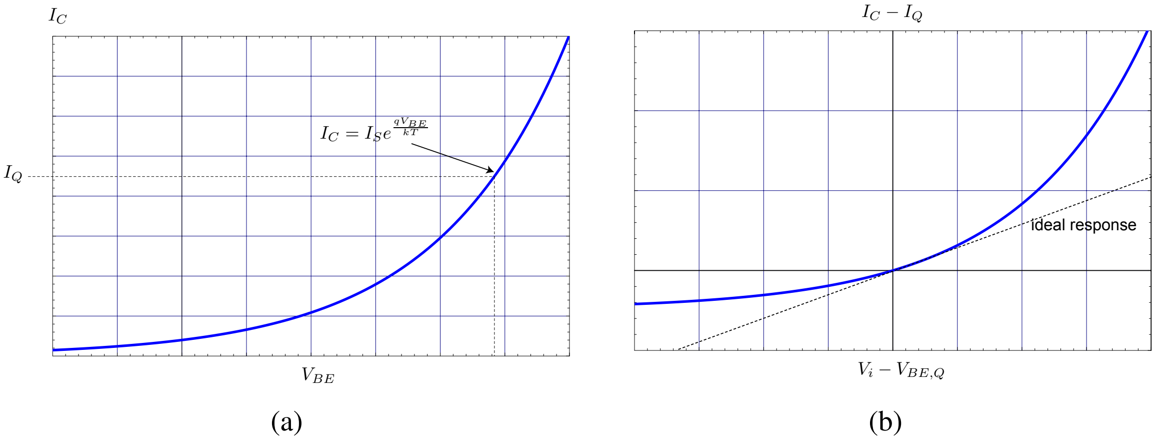


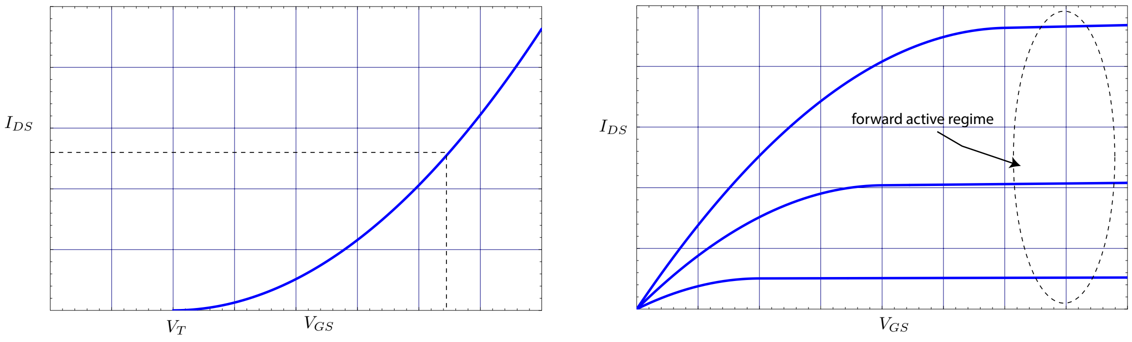
Figure 9.1: The output of a generic amplifier versus the input. If the input is sufficiently large, the output saturates. Only in a small region about the origin is the amplifier "linear".



(a)

(b)

Figure 9.2: (a) The exponential relationship between the collector current and base-emitter voltage of a bipolar transistor. (b) Input/output relation for an AC signal of a BJT amplifier operating around a given bias point.



(a) (b)

Figure 9.3: (a) The square law input/output I_{DS} versus V_{GS} relationship for a MOSFET. (b) The I_{DS} versus V_{DS} relationship for a MOSFET. Amplifiers are biased in the "forward active" regime so the output current is only a strong function of V_{GS}.

curve of a MOSFET in moderate and weak inversion is easy to describe in a "piece-meal" fashion, but difficult to capture with a single equation. Short-channel devices are even more difficult due to velocity saturation and drain induced barrier lowering. We shall return to the topic of MOSFET modeling in a later section.

JFET Distortion

JFETs are useful devices in RF circuits due to their lower noise. The *I*-*V* relation (Fig. 9.4) is also approximately square law but unlike typical MOSFET devices, the "pinch-off" voltage (or threshold voltage) is negative. The JFET current is parameterized by the pinch off voltage V_P and the factor I_{DSS}, or the current for zero V_{GS}

$$I_D = I_{DSS}\left(1 - \frac{V_{GS}}{V_P}\right)^2 \tag{9.4}$$

The gate current (junction leakage) is typically very small $I_G \sim 10^{-12}$A, so for all practical purposes, $R_i = \infty$.

Differential Pair

The differential pair shown in Fig. 9.5a is an important analog and RF building block. For a BJT differential pair, we have $V_i = V_{BE1} - V_{BE2}$

$$I_{C1,2} = I_S e^{\frac{qV_{BE1,2}}{kT}} \tag{9.5}$$

The sum of the collector currents are equal to the current source $I_{C1} + I_{C2} = I_{EE}$. The ideal BJT differential pair *I*-*V* relationship (neglecting base and emitter resistance) is give by

$$I_o = I_{C1} - I_{C2} = \alpha I_{EE} \tanh \frac{qV_i}{2kT} \tag{9.6}$$

This relationship is plotted in Fig. 9.5b. Notice that the output current saturates for large input voltages since the maximum output current is limited to I_{EE}.

9.2 Power Series Relation

For a general circuit, let's represent the memoryless behavior with a power series

$$s_o = f(s_i) \approx a_1 s_i + a_2 s_i^2 + a_3 s_i^3 + \dots \tag{9.7}$$

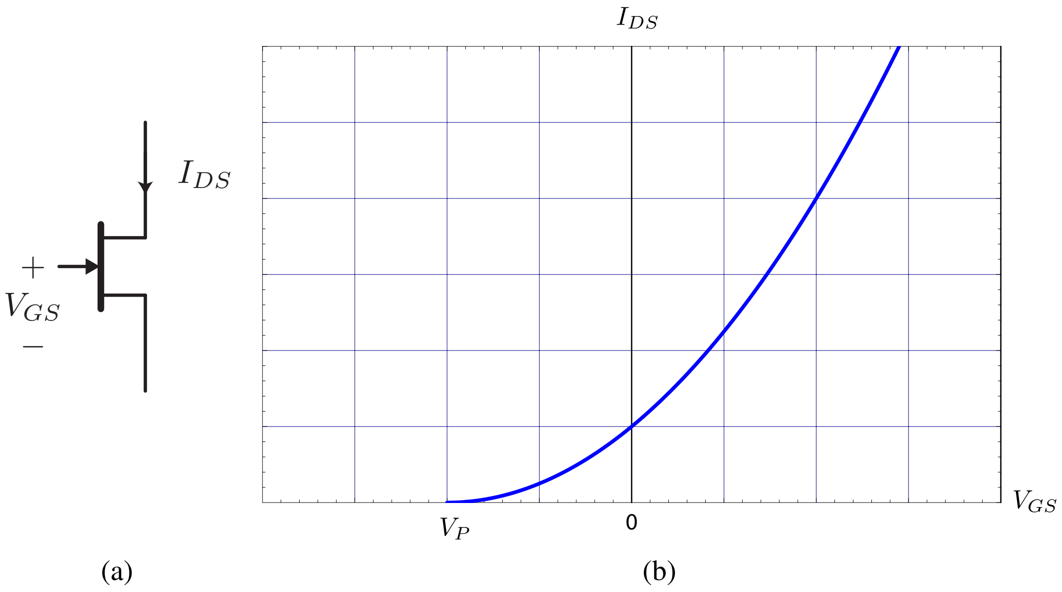


Figure 9.4: The square law relationship for a JFET.

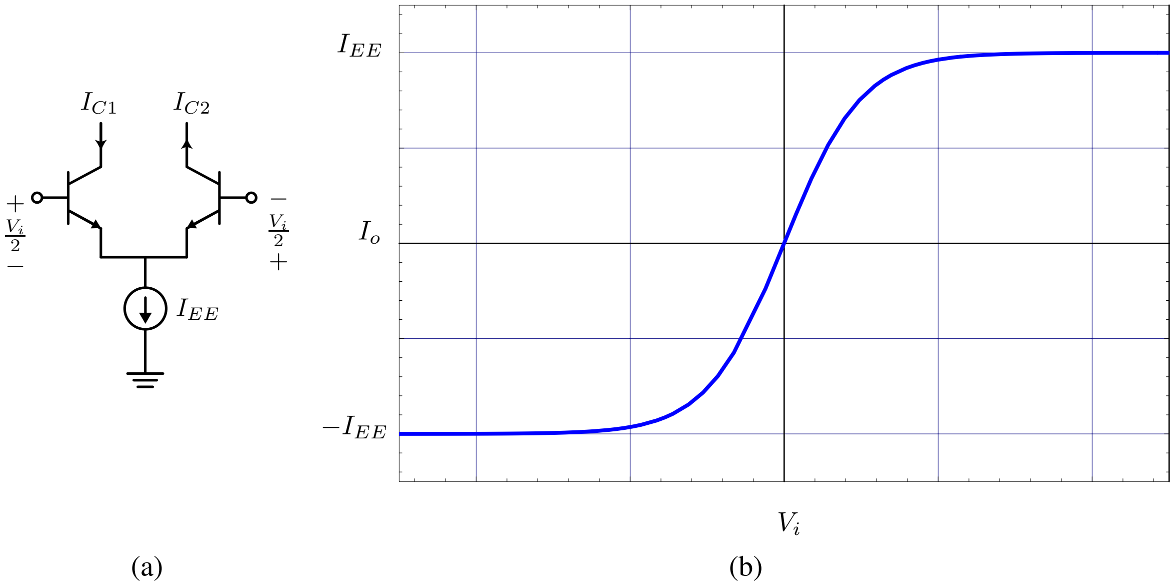


Figure 9.5: (a) A bipolar differential pair driven with a differential input signal. (b) The input/output relationship for a differential pair circuit.

where a_1 is the small signal gain. The coefficients a_1,a_2,a_3,... are independent of the input signal s_i *but* they depend on bias, temperature, and other factors. The coefficients of this power series can be found using Taylor's Theorem

$$a_n = \frac{1}{n!} \frac{d^n f}{d s_i^2} \bigg|_{s_i = Q} \tag{9.8}$$

where the derivatives are evaluated at the quiescent point (Q point). In most scenarios, the DC Q point is subtracted out of the input-output relation, so by definition $a_0 = 0$. We find a_1 by examining the slope

$$a_1 = \frac{1}{1!} \frac{df}{ds_i} \bigg|_{s_i = Q} \tag{9.9}$$

and the term a_2 quadratic term is found by direct application of the equation

$$a_2 = \frac{1}{2!} \frac{df}{ds_i} \bigg|_{s_i = Q} \tag{9.10}$$

For a small signal s_i, we can model the input-output relation with just a few terms, often no more than three, so this procedure is easily carried out by hand.

9.2.1 Harmonic Distortion

Assume we drive the amplifier with a time harmonic signal at frequency ω_1

$$s_i = S_1 \cos \omega_1 t \tag{9.11}$$

A linear amplifier would output $s_o = a_1 S_1 \cos \omega_1 t$ whereas our amplifier generates *distortion*

$$s_o = a_1 S_1 \cos \omega_1 t + a_2 S_1^2 \cos^2 \omega_1 t + a_3 S_1^3 \cos^3 \omega_1 t + \ldots \tag{9.12}$$

or

$$s_o = a_1 S_1 \cos \omega_1 t + \frac{a_2 S_1^2}{2} (1 + \cos 2\omega_1 t) + \frac{a_3 S_1^3}{4} (\cos 3\omega_1 t + 3 \cos \omega_1 t) + \ldots \tag{9.13}$$

The term $a_1 s_1 \cos \omega_1 t$ is the wanted signal, but higher harmonics are also generated. These are usually unwanted and thus called "distortion" terms (there are applications for these harmonics, such as frequency multipliers). We can see that the second-harmonic $\cos 2\omega_1 t$ and third harmonic $\cos 3\omega_1 t$ are generated by the non-linear terms. Also the second order non-linearity produces a DC shift of $\frac{1}{2} a_2 S_1^2$, whereas the third order generates both third-order distortion and more fundamental (or signal at the same frequency as the input). The sign of a_3/a_1 determine whether the distortion product $a_3 S_1^3 \frac{3}{4} \cos \omega_1 t$ adds or subtracts from the fundamental. If the signals add, we say there is gain expansion. If it subtracts, we say there is gain compression.

Waveform Distortion

In Fig. 9.6 we demonstrate the waveform distortion due to second harmonic only. At one peak the second harmonic adds in phase to the fundamental which increases the peak value whereas at the negative peak the signal subtracts, tending to flatten out the waveform. In Fig. 9.7 we show the effects of the third harmonic, where we assume the third harmonic is in phase with the fundamental. Contrast this with Fig. 9.8, where we show the effects of the third harmonic assuming the third harmonic is out of phase with the fundamental. When the signals are in phase, the effect of the third harmonic is to flatten out or "square" the waveform. If out of phase, the third harmonic sharpens the waveform, giving it a triangular appearance.

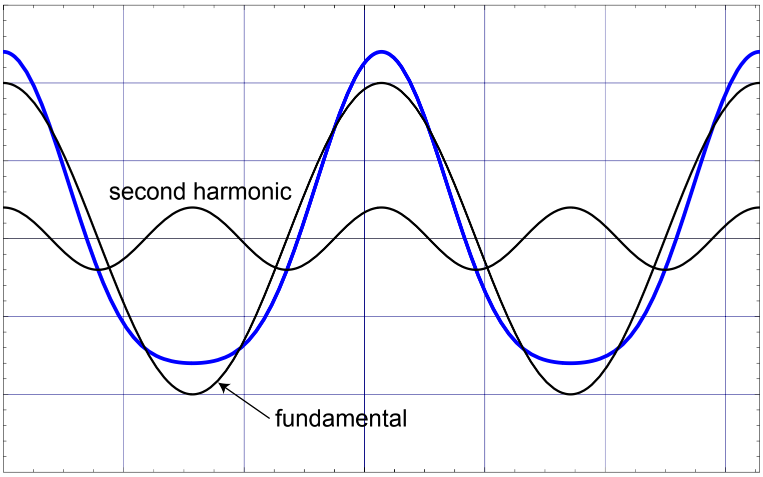


Figure 9.6: A sinusoidal waveform distorted with second harmonic.

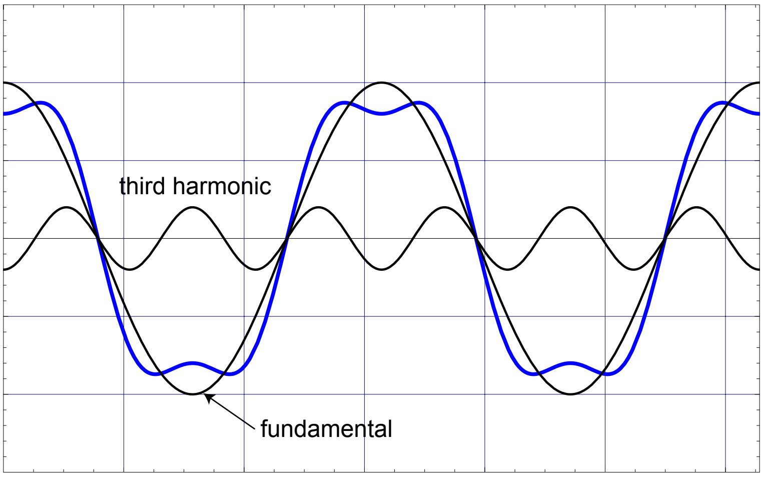


Figure 9.7: A sinusoidal waveform distorted with an in-phase third harmonic.

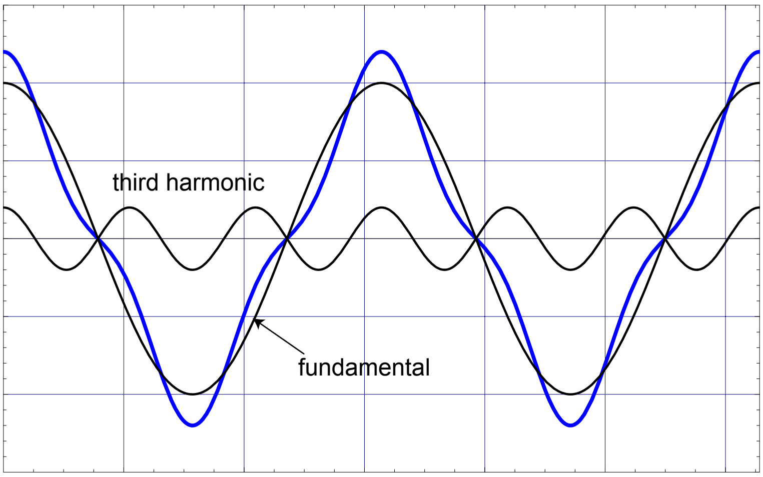


Figure 9.8: A sinusoidal waveform distorted with an out-of-phase third harmonic.

It should be noted that here we have grossly exaggerated the magnitude of the second and third order terms to make the visualization possible. In reality, a good linear amplifier produces so little second and third order that the waveforms look essentially sinusoidal to the naked eye. To see the impact of the distortion, we must look at the signal spectrum on a log scale, where the distortion products are clearly visible.

9.2.2 General Distortion Term

To understand the distortion generation for the n'th order power, use Euler's Theorem and write $\cos^n\theta = \frac{1}{2^n}\left(e^{j\theta}+e^{-j\theta}\right)^n$. This allows us to use Binomial formula to expand the product

$$\cos^n\theta = \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}e^{jk\theta}e^{-j(n-k)\theta} \tag{9.14}$$

Take for example, for $n=3$ we have

$$= \frac{1}{8}\left(\binom{3}{0}e^{-j3\theta}+\binom{3}{1}e^{j\theta}e^{-j2\theta}+\binom{3}{2}e^{j2\theta}e^{-j\theta}+\binom{3}{3}e^{j3\theta}\right) \tag{9.15}$$

$$= \frac{1}{8}\left(e^{-j3\theta}+e^{j3\theta}\right)+\frac{1}{8}3\left(e^{j\theta}+e^{-j\theta}\right) = \frac{1}{4}\cos 3\theta+\frac{3}{4}\cos\theta \tag{9.16}$$

Using this particular example, we notice a few trends. For example, an odd power n, we will see a pairing up of positive and negative powers of exponentials. On the other hand, if n is an even power, the middle term is the unpaired DC term

$$\binom{2k}{k}e^{jk\theta}e^{-jk\theta} = \binom{2k}{k} \tag{9.17}$$

We conclude that in general only even powers in the transfer function can shift the DC operation point. Return now to the general term in the binomial expansion, consider that $(x+x^{-1})^n$ is given by

$$\binom{n}{k}x^{n-k}x^{-k} = \binom{n}{k}x^{n-2k} \tag{9.18}$$

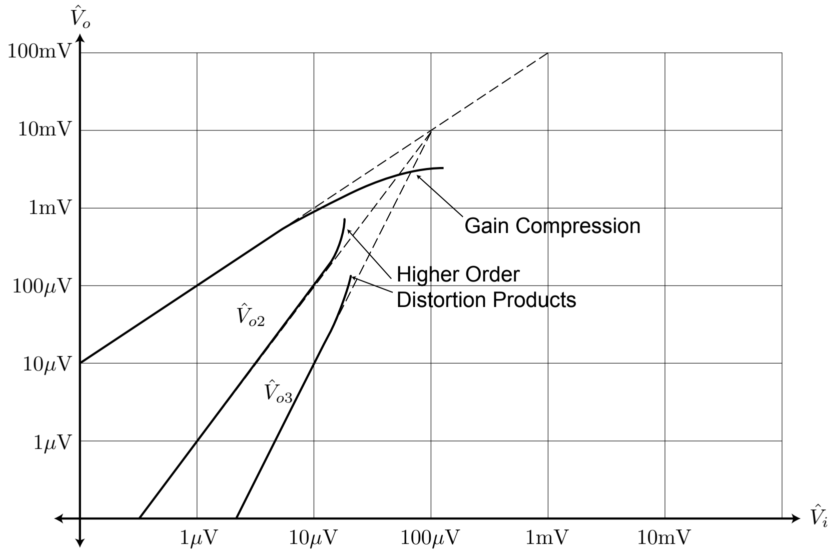


Figure 9.9: A plot of the input/output relations for the fundamental, second, and third harmonic components of the output signal.

From here we see that the term $\binom{n}{k}x^{n-2k}$ generates every other harmonic. If n is even, then only even harmonics are generated (including DC). If n is odd, only odd harmonics are generated. These observations turn out to be very important as we will show later.

It's interesting to observe that certain functions have even or odd symmetry. Recall that an "odd" function $f(-x) = -f(x)$ (anti-symmetric) has an odd power series expansion

$$f(x) = a_1 x + a_3 x^3 + a_5 x^5 + \dots \tag{9.19}$$

Whereas an even function, $g(-x) = g(x)$, has an even power series expansion

$$g(x) = a_0 + a_2 x^2 + a_4 x^4 + \dots \tag{9.20}$$

9.2.3 Harmonic Distortion Metrics

From the above discussion, we see that for a sinusoidal input, the output of system with memoryless non-linear is new tones at various harmonics of the input frequency. So in general, for a sinusoidal excitation, the output of a memoryless non-linearity can be written as

$$v_o = \hat{V}_{o1}\cos\omega_1 t + \hat{V}_{o2}\cos 2\omega_1 t + \hat{V}_{o3}\cos 3\omega_1 t + \dots \tag{9.21}$$

If we plot the various harmonics as a function of the input level, we generate a plot similar to Fig. 9.9 (note the small-signal gain is 100). We will now define metrics that measure the spectral purity of the output of the amplifier. For example, we can measure the relative amount of second or third harmonic content in the output waveform.

The fractional second-harmonic distortion is a commonly cited metric and defined as

$$HD_2 = \frac{\text{ampl of second harmonic}}{\text{ampl of fund}} \tag{9.22}$$

Note that all even powers generate second harmonic distortion, but here we assume the a_2 term dominates. This is a good assumption as long as the input signal is small, in other words if the signal does not deviate significantly from the quiescent operating point. So assuming that the square power dominates the second-harmonic,

$$HD_2 = \frac{a_2 \frac{S_1^2}{2}}{a_1 S_1} \tag{9.23}$$

or

$$HD_2 = \frac{1}{2}\frac{a_2}{a_1} S_1 \tag{9.24}$$

In a similar way, the fractional third harmonic distortion is defined as

$$HD_3 = \frac{\text{ampl of third harmonic}}{\text{ampl of fund}} \tag{9.25}$$

If we assume that the cubic power dominates the third harmonic

$$HD_3 = \frac{a_3 \frac{S_1^2}{4}}{a_1 S_1} \tag{9.26}$$

or

$$HD_3 = \frac{1}{4}\frac{a_3}{a_1} S_1^2 \tag{9.27}$$

Output Referred Harmonic Distortion

In terms of the output signal S_{om}, if we again neglect gain expansion/compression, we have $S_{om} = a_1 S_1$. This allows us to state the HD distortion metrics in terms of the output signal amplitude

$$HD_2 = \frac{1}{2}\frac{a_2}{a_1^2} S_{om} \tag{9.28}$$

$$HD_3 = \frac{1}{4}\frac{a_3}{a_1^3} S_{om}^2 \tag{9.29}$$

We may conclude that on a dB scale, as we vary the input power, the second harmonic increases linearly with a slope of one whereas the third harmonic increases with a slope of two.

Signal Power

Notice that the output signal for a sinusoidal input is in a Fourier series

$$v_o(t) = \hat{V}_{o1} \cos \omega_1 t + \hat{V}_{o2} \cos 2\omega_1 t + \hat{V}_{o3} \cos 3\omega_1 t + \ldots \tag{9.30}$$

By Parseval's theorem, we know the total power in the signal is related to the power in the harmonics

$$\int_T v^2(t)dt = \int_T \sum_j \hat{V}_{oj} \cos(j\omega_1 t) \sum_k \hat{V}_{ok} \cos(k\omega_1 t)dt \tag{9.31}$$

$$= \sum_j \sum_k \int_T \hat{V}_{oj} \cos(j\omega_1 t) \hat{V}_{ok} \cos(k\omega_1 t)dt \tag{9.32}$$

By the orthogonality of the harmonics, we obtain Parseval's Theorem

$$\int_T v^2(t)dt = \sum_j \sum_k \tfrac{1}{2}\delta_{jk}\hat{V}_{oj}\hat{V}_{ok} = \tfrac{1}{2}\sum_j \hat{V}_{oj}^2 \tag{9.33}$$

The power in the distortion relative to the fundamental power is therefore given by

$$\frac{\text{Power in Distortion}}{\text{Power in Fundamental}} = \frac{V_{o2}^2}{V_{o1}^2} + \frac{V_{o3}^2}{V_{o1}^2} + \cdots \tag{9.34}$$

$$= HD_2^2 + HD_3^2 + HD_4^2 + \cdots \tag{9.35}$$

Total Harmonic Distortion

Motivated by the previous results, we define the *Total Harmonic Distortion* (THD) by the following expression

$$THD = \sqrt{HD_2^2 + HD_3^2 + \cdots} \tag{9.36}$$

Based on the particular application, we specify the maximum tolerable THD. Telephone audio can be pretty distorted and $THD < 10\%$ is acceptable. High quality audio is very sensitive, $THD < 1\%$ to $THD < .001\%$. Video is also pretty forgiving, $THD < 5\%$ for most applications. Analog repeaters require very low distortion or $THD < .001\%$. For RF amplifiers $< 0.1\%$. These numbers are general guidelines. Each particular application will dictate a required specification.

Example 16:

Distortion of BJT Amplifiers

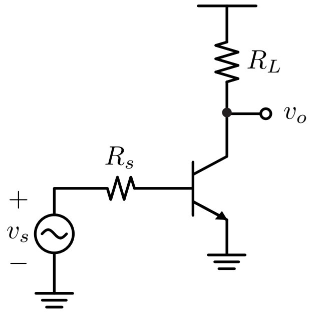


Figure 9.10: A common-emitter (CE) bipolar amplifier.

Consider the CE BJT amplifier shown in Fig. 9.10. The biasing is omitted for clarity. The output voltage is simply

$$V_o = V_{CC} - I_C R_C \tag{9.37}$$

Therefore the distortion is generated by I_C alone. Recall that the collector current is an exponential function of the base-emitter voltage

$$I_C = I_S e^{qV_{BE}/kT} \tag{9.38}$$

Now assume the input $V_{BE} = v_i + V_Q$, where V_Q is the bias point. The current is therefore given by

$$I_C = \underbrace{I_S e^{\frac{V_Q}{V_T}}}_{I_Q} e^{\frac{v_i}{V_T}} \tag{9.39}$$

Using a Taylor expansion for the exponential

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots \tag{9.40}$$

$$I_C = I_Q(1 + \frac{v_i}{V_T} + \frac{1}{2}\left(\frac{v_i}{V_T}\right)^2 + \frac{1}{6}\left(\frac{v_i}{V_T}\right)^3 + \cdots) \tag{9.41}$$

Define the output AC signal as $i_c = I_C - I_Q$,

$$i_c = \frac{I_Q}{V_T}v_i + \frac{1}{2}\left(\frac{q}{kT}\right)^2 I_Q v_i^2 + \frac{1}{6}\left(\frac{q}{kT}\right)^3 I_Q v_i^3 + \cdots \tag{9.42}$$

Compare to $S_o = a_1 S_i + a_2 S_i^2 + a_3 S_i^3 + \cdots$

$$a_1 = \frac{qI_Q}{kT} = g_m \tag{9.43}$$

$$a_2 = \frac{1}{2}\left(\frac{q}{kT}\right)^2 I_Q \tag{9.44}$$

$$a_3 = \frac{1}{6}\left(\frac{q}{kT}\right)^3 I_Q \tag{9.45}$$

Note that the linear term is just the transistor small-signal transconductance. This is a good check on our calculations.

For any BJT or HBT (Si, SiGe, Ge, GaAs) device, we have the following result

$$HD_2 = \frac{1}{4}\frac{q\hat{v}_i}{kT} \tag{9.46}$$

where $\hat{v}_i$ is the peak value of the input sine voltage. For $\hat{v}_i = 10\text{mV}$, $HD_2 = 0.1 = 10\%$. This is a surprising result which implies that an amplifier like this in open-loop configuration should only be used to process very small signals (microvolts to millivolts), otherwise the output waveform will be heavily distorted.

We can also express the distortion as a function of the output current swing $\hat{i_c}$

$$HD_2 = \frac{1}{2}\frac{a_2}{a_1^2}S_{om} = \frac{1}{4}\frac{\hat{i_c}}{I_Q} \tag{9.47}$$

For $\frac{\hat{i_c}}{I_Q} = 0.4$, $HD_2 = 10\%$. For example in a power amplifier or driver application, we may have to increase the bias current I_Q beyond the limits imposed by bandwidth or drive strength simply in order to meet the distortion requirements.

9.2.4 Intermodulation Distortion

So far we have characterized a non-linear system for a single tone. What if we apply more than one tone? Due to non-linearity, we cannot use the Superposition Theorem. Instead, we have to consider each case separately. Let's begin with two tones

$$S_i = S_1 \cos\omega_1 t + S_2 \cos\omega_2 t \tag{9.48}$$

and assume the system is described by a power series

$$S_o = a_1 S_i + a_2 S_i^2 + a_3 S_i^3 + \cdots \tag{9.49}$$

Substituting the input signal we have

$$= a_1 S_1 \cos\omega_1 t + a_1 S_2 \cos\omega_2 t + a_3 (S_i)^3 + \cdots \tag{9.50}$$

Focus on the second power term:

$$a_2 S_1^2 \cos^2\omega_1 t + a_2 S_2^2 \cos^2\omega_2 t + 2a_2 S_1 S_2 \cos\omega_1 t \cos\omega_2 t \tag{9.51}$$

or simplifying,

$$= a_2 \frac{S_1^2}{2}(\cos 2\omega_1 t + 1) + a_2 \frac{S_2^2}{2}(\cos 2\omega_2 t + 1) + a_2 S_1 S_2 \left(\cos(\omega_1 + \omega_2)t - \cos(\omega_1 - \omega_2)t\right) \tag{9.52}$$

Second Order Intermodulation

The last term $\cos(\omega_1 \pm \omega_2)t$ is the second-order intermodulation term. Note that this term arises due to the non-linearity and is not predicted from our single tone analysis. It's a new frequency that is not a harmonic of either input signal, but rather an intermodulation of the two frequencies. The intermodulation distortion metric IM_2 is defined when the two input signals have equal amplitude $S_i = S_1 = S_2$

$$IM_2 = \frac{\text{Amp of Intermod}}{\text{Amp of Fund}} = \frac{a_2}{a_1} S_i \tag{9.53}$$

From our previous calculation, we note the relation between IM_2 and HD_2

$$IM_2 = 2HD_2 = HD_2 + 6\text{dB} \tag{9.54}$$

This term produces distortion at a lower frequency $\omega_1 - \omega_2$ and at a higher frequency $\omega_1 + \omega_2$. In theory, for a narrowband system, we could use filters to attenuate these distortion products.

Example 17:

Consider a wideband Software Defined Radio (SDR) receiver with an input bandwidth from 800MHz – 2.4GHz. Suppose that two unwanted interfering signals appear at 800MHz and 900MHz. Then we see that the second-order distortion will produce distortion at 100MHz and 1.7GHz. Since 1.7GHz is in the receiver band, signals at this frequency will be corrupted by the distortion. A weak signal in this band can be "swamped" by the distortion. In this example, we cannot simply filter out the distortion product since the receiver is wideband.

Apparently, a "narrowband" system does not suffer from IM_2? Or does it ?

Example 18:Low-IF Receiver

In a low-IF or direct conversion receiver, the signal is down-converted to a low intermediate frequency f_{IF}. Since $\omega_1 - \omega_2$ can potentially produce distortion at low frequency, IM_2 is very important in such systems.

Let's take an example of a narrowband system has a receiver bandwidth of 1.9GHz - 2.0GHz. A sharp input filter eliminates any interference outside of this band. The IF frequency is 1MHz. Imagine two interfering signals that appear at $f_1 = 1.910$GHz and $f_2 = 1.911$GHz. Notice that $f_2 - f_1 = f_{IF}$. Thus the output of the amplifier/mixer generate distortion at the IF frequency, potentially disrupting the communication.

Cubic Itermodulation

So far we have applied two tones to our system and focused only on the second order terms. Now let's consider the output of the cubic term

$$a_3 s_i^3 = a_3(S_1 \cos\omega_1 t + S_2 \cos\omega_2 t)^3 \tag{9.55}$$

First notice that the first and last term in the expansion are the same as the cubic distortion with a single input

$$\frac{a_3 S_{1,2}^3}{4}(\cos 3\omega_{1,2}t + 3\cos\omega_{1,2}t) \tag{9.56}$$

The cross product terms look like

$$\binom{3}{2} a_3 S_1 S_2^2 \cos\omega_1 t \cos^2\omega_2 t \tag{9.57}$$

which can be simplified to

$$3\cos\omega_1 t\cos^2\omega_2 t = \frac{3}{2}\cos\omega_1 t(1+\cos 2\omega_2 t) = \tag{9.58}$$

$$= \frac{3}{2}\cos\omega_1 t + \frac{3}{4}\cos(2\omega_2 \pm \omega_1) \tag{9.59}$$

The interesting term is the intermodulation at $2\omega_2 \pm \omega_1$. By symmetry, then, we also generate a term like

$$a_3 S_1^2 S_2 \frac{3}{4}\cos(2\omega_1 \pm \omega_2) \tag{9.60}$$

Similar to the second-order intermodulation terms, these third order intermodulations only appear when we apply more than one tone and are not at integer harmonics of the input frequencies.

Cubic distortion plays a central role in communication systems because of the following property. If $\omega_1 \approx \omega_2$, then the intermodulation $2\omega_2 - \omega_1 \approx \omega_1$. This means that we can never filter these distortion products because they fall in-band.

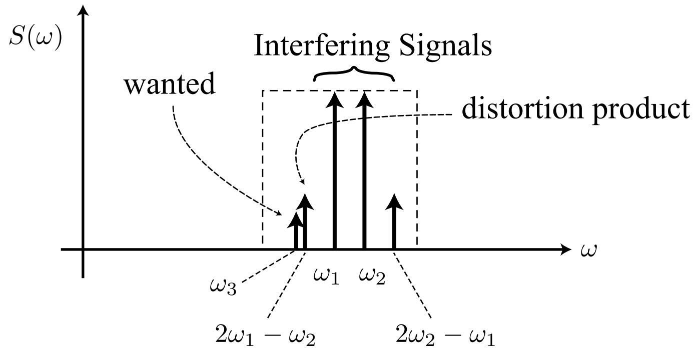


Figure 9.11: Suppose that a weak signal must be detected in the presence of two in-band strong tones. The IM_3 distortion generated by two tones falls very close to a desired signal at ω_3 within the band of the system.

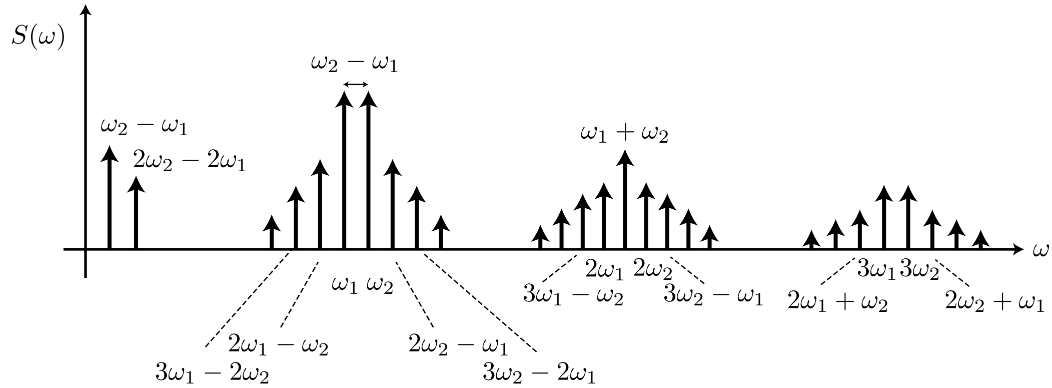


Figure 9.12: The output spectrum of a memoryless amplifier excited by two-tones contains many harmonic and intermodulation components.

In-Band IM3 Distortion

Now we see that even if the system is narrowband, the output of an amplifier can contain in band intermodulation due to IM_3, shown in Fig. 9.11. This is in contrast to IM_2 where the frequency of the intermodulation was at a lower and higher frequency. The IM_3 frequency can fall in-band for two in-band interferers (such as other users of the spectrum).

We define IM_3 in a similar manner for $S_i = S_1 = S_2$

$$IM_3 = \frac{\text{Amp of Third Intermod}}{\text{Amp of Fund}} = \frac{3}{4}\frac{a_3}{a_1}S_i^2 \tag{9.61}$$

Note the relation between IM_3 and HD_3

$$IM_3 = 3HD_3 = HD_3 + 10\text{dB} \tag{9.62}$$

Complete Two-Tone Response

We have so far identified the harmonics and IM_2 and IM_3 products. A more detailed analysis shows that an order n non-linearity can produce intermodulation at frequencies $j\omega_1 \pm k\omega_2$ where $j + k = n$. As shown in Fig. 9.12, all tones are spaced by the difference $\omega_2 - \omega_1$. We will return to this result and show that this is true.

Example 19:

BJT IM3

Using the results from the previous example, let's see the maximum allowed signal for $IM_3 \leq 1\%$

$$IM_3 = \frac{3}{4}\frac{a_3}{a_1}S_1^2 = \frac{1}{8}\left(\frac{q\hat{v}_i}{kT}\right)^2 \tag{9.63}$$

Solve $\hat{v}_i = 7.3\text{mV}$. That's a pretty small voltage and for many practical applications we'd like to improve the linearity of this amplifier. Later on we'll explore feedback techniques that improve the linearity.

Example 20:

MOSFET Distortion

Consider the simple single stage MOS common-source amplifier shown in Fig. 9.13.

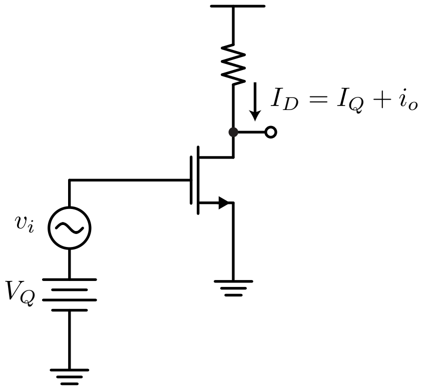


Figure 9.13: A common source MOS amplifier.

For a long channel device, the drain current of the device is given by the square law relation

$$I_D = \frac{1}{2}\mu C_{ox}\frac{W}{L}(V_{GS} - V_T)^2 \tag{9.64}$$

writing the current in terms of an AC and DC component

$$i_o + I_Q = \frac{1}{2}\mu C_{ox}\frac{W}{L}(V_Q + v_i - V_T)^2 \tag{9.65}$$

Expanding the square

$$= \frac{1}{2}\mu C_{ox}\frac{W}{L}\left\{(V_Q - V_T)^2 + v_i^2 + 2v_i(V_Q - V_T)\right\} \tag{9.66}$$

which allows us to separate the DC And AC terms

$$= \underbrace{I_Q}_{\text{dc}} + \underbrace{\mu C_{ox}\frac{W}{L}v_i(V_Q - V_T)}_{\text{linear}} + \underbrace{\tfrac{1}{2}\mu C_{ox}\frac{W}{L}v_i^2}_{\text{quadratic}} \tag{9.67}$$

Clearly, an ideal square law device only generates 2nd order distortion

$$i_o = g_m v_i + \frac{1}{2}\mu C_{ox}\frac{W}{L}v_i^2 \tag{9.68}$$

$$a_1 = g_m \tag{9.69}$$

$$a_2 = \frac{1}{2}\mu C_{ox}\frac{W}{L} = \frac{1}{2}\frac{g_m}{V_Q - V_T} \tag{9.70}$$

$$a_3 \equiv 0 \tag{9.71}$$

The harmonic distortion is given by

$$HD_2 = \frac{1}{2}\frac{a_2}{a_1}v_i = \frac{1}{4}\frac{g_m}{V_Q - V_T}\frac{1}{g_m}v_i = \frac{1}{4}\frac{v_i}{V_Q - V_T} \tag{9.72}$$

$$HD_3 = 0 \tag{9.73}$$

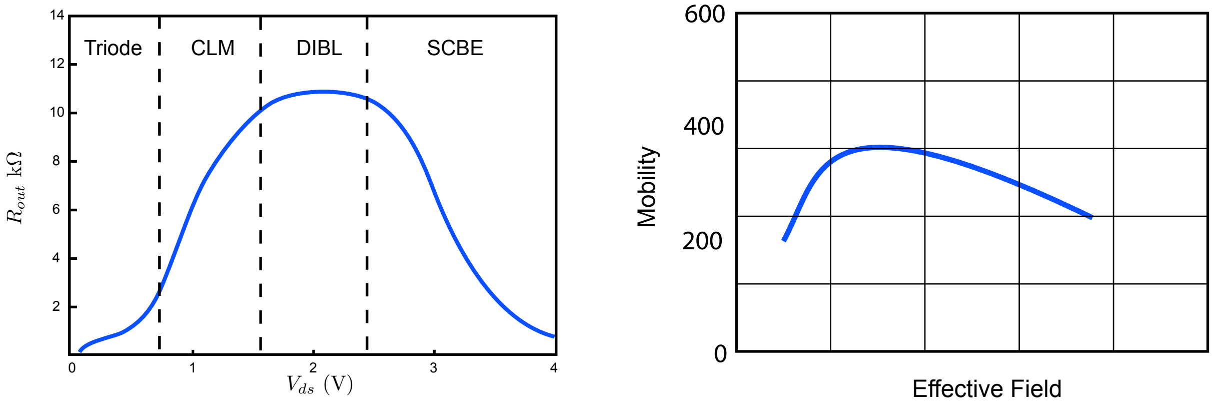


Figure 9.14: (a) The output resistance of a short-channel MOSFET in different regions of operation (after BSIM4 manual). (b) The mobility of a MOSFET device as a function of vertical electrical field (controlled by V_{GS}).

We would therefore expect that a MOS amplifier should not generate any IM_3 products, which are key in-band distortion products for narrowband RF applications. Unfortunately, the real MOSFET device generates higher order distortion through several additional mechanisms.

As shown in Fig. 9.14a, the output impedance is non-linear, which means that unless the load is much smaller than the device output resistance, cubic distortion will be generated. More importantly, the mobility μ is not a constant but a function of the vertical and horizontal electric field (Fig. 9.14b. Including this effect also introduces cubic and higher terms in the MOS I-V curve. More importantly, if we bias the device in moderate or weak inversion, where the device behavior is more exponential and therefore more non-linear, cubic distortion is generated similar to a BJT. Finally, we'll learn that while feedback can linearize an amplifier, it can also generate cubic distortion from a purely square law device. In a MOSFET, there is internal *feedback* due to source resistance R_S and feedback capacitance C_{gd}, and so third order is always observed.

9.3 System Level Specifications

9.3.1 Definition of Gain

When the input signal is small, gain is easy to define and most definitions tend to agree. A common definiton for small-signal gain is to evaluate the slope of the input-output relationship of the amplifier around the operating point of inerest. But if the signal excursion is large, then the amplifier leaves the linear regime and the effective amplitude changes, as we have observed. As shown in Fig. 9.15a, the small-signal gain is related to the slope of the input-output curve at a given point. Since the curve flattens out when the amplifier "rails", the small-signal gain (slope) goes to zero.

Since the output is going to be rich in harmonics, we could focus our attention on the fundamental frequency of interest and define gain as follows

$$G = \frac{V_{o,\omega_0}}{V_i} \tag{9.74}$$

where we assume the input is a sinusoid at frequency ω_0 and the output funamental harmonic amplitude is defined as the desired output signal. Now we can see that as the amplifier slope changes as a function of the input drive, it's not so clear how to define gain. For example, consider the limit as the amplifier is driven so hard that the output clips to the positive and negative rails. Then we have

$$V_o = V_{rail} \text{sign}\left(\sin(\omega_0 t)\right) \tag{9.75}$$

or a square wave at the output. If we take the amplitude of the fundamental, we have

$$V_{o,\omega_0} = \frac{4}{\pi} V_{rail} \tag{9.76}$$

where $4/\pi$ is the amplitude of the first harmonic of the Fourier series of a ± 1 square wave. The gain therefore reduces with the input drive

$$G = \frac{4}{\pi} \frac{V_{rail}}{V_i} \tag{9.77}$$

or the gain tends to compress. Interestingly, before this happens, when the amplifier is only driven with a signal to produce a weakly non-linear response, we can also observe deviations in the gain from the small-signal value.

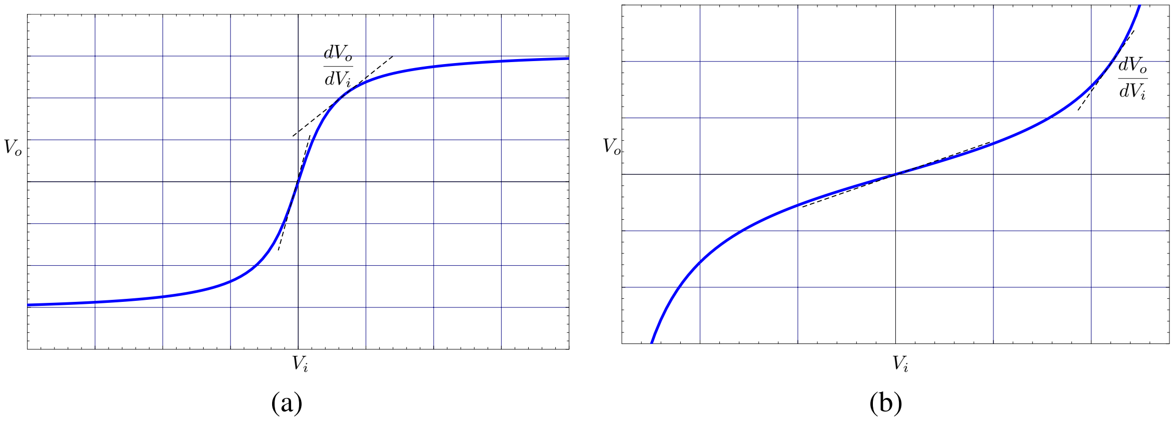


Figure 9.15: (a) An amplifier input-output response exhibiting (a) gain compression and (b) gain expansion.

9.3.2 Gain Compression

The large signal input/output relation can display gain compression or expansion, or possibly both. In Fig. 9.15b, we show an amplifier with gain expansion. In practice the range of gain expansion is limited by physical constraints such as supply voltage or thermal effects. Even if the power supply were increased, eventually the current through the active devices would melt the wires!

1-dB Compression Point

Due to gain compression, eventually the output signal drops and so if plot the gain (log scale) as a function of the input power, we identify the point where the gain has dropped by 1 dB as the 1 dB compression point. In Fig. 9.16 we show a hypothetical amplifier and the 1 dB compression point. As we shall see, this is a very important number to keep in mind since most amplifiers act “linearly” up to this point.

Apparent Gain

Recall that around a small deviation, the large signal curve is described by a polynomial

$$s_o = a_1 s_i + a_2 s_i^2 + a_3 s_i^3 + \cdots \tag{9.78}$$

For an input $s_i = S_1 \cos(\omega_1 t)$, the cubic term generates

$$S_1^3 \cos^3(\omega_1 t) = S_1^3 \cos(\omega_1 t) \frac{1}{2}(1 + \cos(2\omega_1 t)) \tag{9.79}$$

$$= S_1^3 \left(\frac{1}{2} \cos(\omega_1 t) + \frac{2}{4} \cos(\omega_1 t) \cos(2\omega_1 t) \right) \tag{9.80}$$

Recall that $2\cos a \cos b = \cos(a+b) + \cos(a-b)$

$$= S_1^3 \left(\frac{1}{2} \cos(\omega_1 t) + \frac{1}{4} (\cos(\omega_1 t) + \cos(3\omega_1 t)) \right) \tag{9.81}$$

Collecting terms

$$= S_1^3 \left(\frac{3}{4} \cos(\omega_1 t) + \frac{1}{4} \cos(3\omega_1 t) \right) \tag{9.82}$$

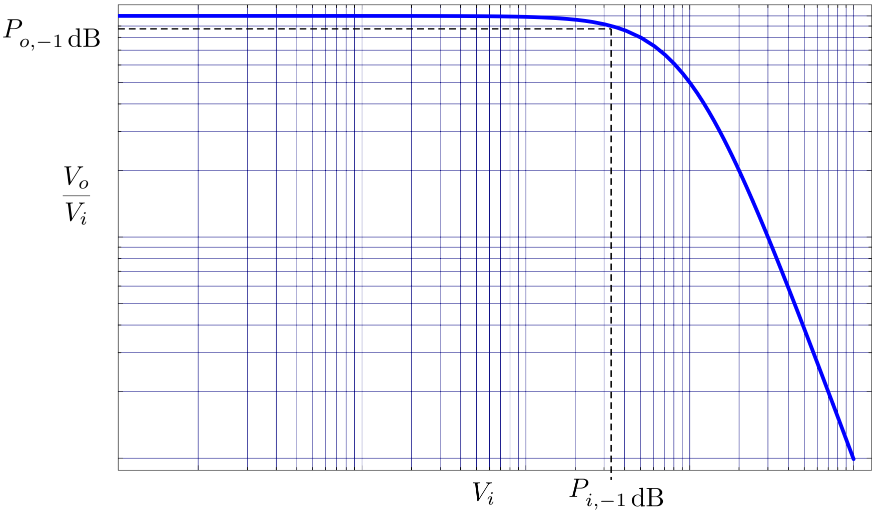


Figure 9.16: The gain of an amplifier plotted versus input signal. When the input power reaches the $P_{i,-1\,\mathrm{dB}}$, the gain drops by 1 dB.

The apparent gain of the system is therefore

$$G = \frac{S_{o,\omega_1}}{S_{i,\omega_1}} = \frac{a_1 S_1 + \frac{3}{4} a_3 S_1^3}{S_1} \tag{9.83}$$

$$= a_1 + \frac{3}{4} a_3 S_1^2 = a_1 \left(1 + \frac{3}{4} \frac{a_3}{a_1} S_1^2\right) = G(S_1) \tag{9.84}$$

If $a_3/a_1 < 0$, the gain compresses with increasing amplitude, otherwise the gain expands.

Calculation of the 1-dB Compression Point

Let's find the input level where the gain has dropped by 1 dB

$$20 \log \left(1 + \frac{3}{4} \frac{a_3}{a_1} S_1^2\right) = -1\,\mathrm{dB} \tag{9.85}$$

$$\frac{3}{4} \frac{a_3}{a_1} S_1^2 = -0.11 \tag{9.86}$$

$$S_1 = \sqrt{\frac{4}{3} \left|\frac{a_1}{a_3}\right|} \times \sqrt{0.11} = IIP3 - 9.6\,\mathrm{dB} \tag{9.87}$$

The term in the square root is called the third-order intercept point (see next few slides).

9.3.3 Intermodulation Intercept Point

Intercept Point IP_2

As shown in Fig. 9.17, the extrapolated point where $IM_2 = 0\,\mathrm{dBc}$ is known as the second order intercept point IP_2. Since the second order IM distortion products increase like s_i^2, we expect that at

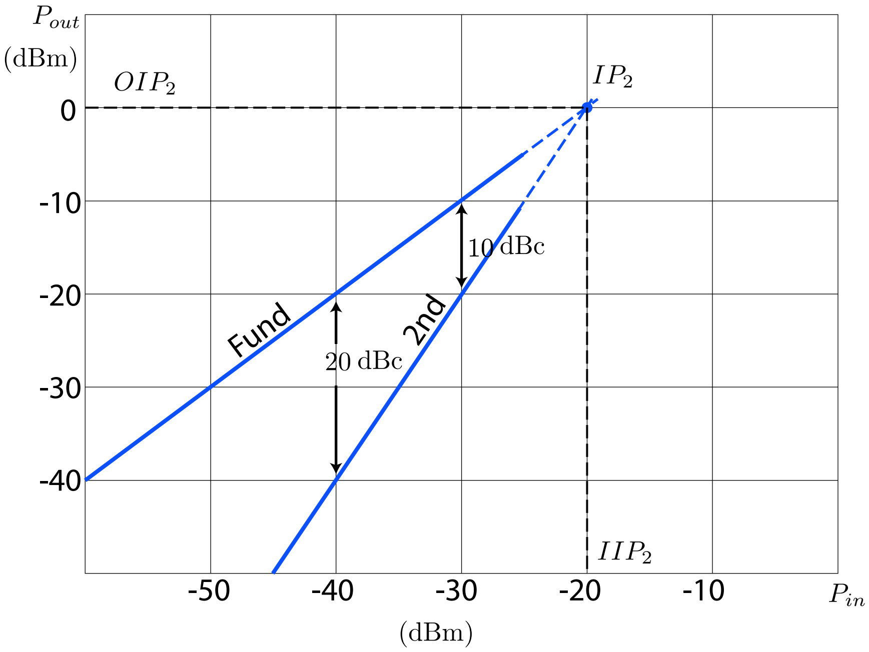


Figure 9.17: A plot of the fundamental and quadratic intermodulation product power versus input power.

some power level the distortion products will overtake the fundamental signal. The extrapolated point where the curves of the fundamental signal and second order distortion product signal meet is the *intercept point* (IP_2). At this point, by definition $IM_2 = 0$ dBc. The input power level when $IM_2 = 0$dBc is known as IIP_2, and the output power when this occurs is the OIP_2 point. Once the IP_2 point is known, the IM_2 at any other power level can be calculated since for a 1 dB back-off from the IP_2 point, the IM_2 improves 1 dB (dB for dB). This is somewhat confusing since the IP_2 products increase 2 dB per every 1 dB of input power, but keep in mind that IM_2 is the ratio (or distance between the curves on a log-scale), which only increases by 1 dB.

One final point about IP_2 is that this is the extrapolanted point, not the actual point measured in the lab. This may seem puzzling at first but the reason that we measure IP_2 at low powers is that we want only second-order distortion terms to be due to a_2, and not higher even powers. This fixes the slope of the IM_2 versus power at the 1-dB value, which makes it convenient to predict IM_2 at other power levels. So a good rule of thumb is to measure IM_2 slope in the lab (or with simulation) and verify the slope is only unity, and then proceed to make the extrapolation.

Intercept Point IP_3

Naturally, the extrapolated point where $IM_3 = 0$ dBc is known as the third-order intercept point IP_3, as shown in Fig. 9.18. Since the third order IM distortion products increase like s_i^3, we expect that at some power level the third-order distortion products will overtake the fundamental signal. The extrapolated point where the curves of the fundamental signal and third order distortion product signal meet is the *intercept point* (IP_3). At this point, by definition $IM_3 = 0$ dBc. This input power level is known as IIP_3, and the output power when this occurs is the OIP_3 point. Once the IP_3 point is known, the IM_3 at any other power level can be calculated. Note that for a 10 dB back-off from the IP_3 point, the IM_3 improves 20 dB. The 2 : 1 slope comes about since the input power increases linearly whereas the cubic products increases with a slope of 3, so the ratio (or distance on the log-scale) has a slope of 2.

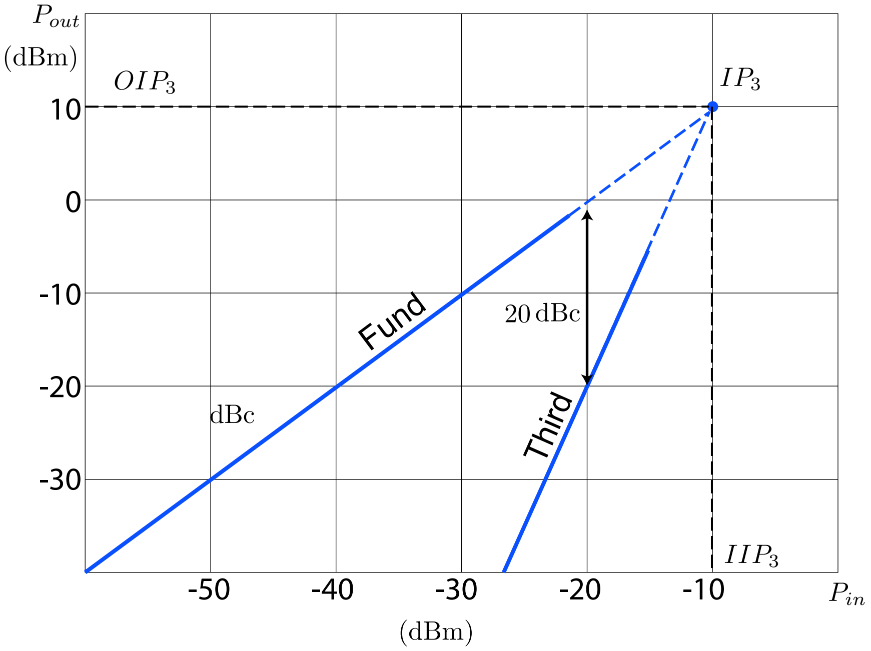


Figure 9.18: A plot of the fundamental and cubic intermodulation product power versus input power.

Example 21:

Intercept Point Example

From the graph of Fig. 9.18 we see that our amplifier has an $IIP_3 = -10\,\text{dBm}$. (a) What's the IM_3 for an input power of $P_{in} = -20\,\text{dBm}$? (b) What's the IM_3 for an input power of $P_{in} = -110\,\text{dBm}$?

(a) Since the IM_3 improves by 20dB for every 10dB back-off, it's clear that $IM_3 = 20\,\text{dBc}$. In this case we can also just read from the graph.

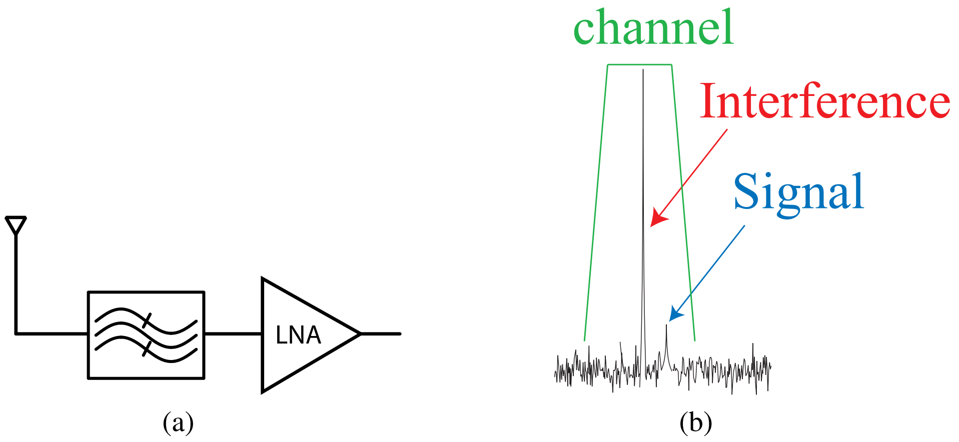
(b) In this case, we don't have measured data, so we rely on the 2:1 slope of IM_3. For -110 dBm, we are backing off from the IP_3 point by 100 dB, so IM_3 should improve by twice that amount or $IM_3 = 200\,\text{dBc}$.

From the above results we see that for small input signals, the amplifier is linear for all intents and purposes, since the distortion products are extremely small and buried in the noise of the amplifier.

Calculated $IIP2/IIP3$

We can also calculate the IIP points directly from our power series expansion. By definition, the $IIP2$ point occurs when

$$IM_2 = 1 = \frac{a_2}{a_1} S_i \tag{9.88}$$



(a) (b)

Figure 9.19: (a) The front-end of every communication system includes a channel filter (to attenuate out of band blockers) and an LNA to improve the sensitivity of the system. (b) In a typical communication system, the desired signal is often accompanied by a larger number of interference signals (or “blockers”), which are simply signals originating from other users of the channel.

Solving for the input signal level

$$IIP_2 = S_i = \frac{a_1}{a_2} \tag{9.89}$$

In a like manner, we can calculate IIP_3

$$IM_3 = 1 = \frac{3}{4}\frac{a_3}{a_1}S_i^2 \tag{9.90}$$

$$IIP_3 = S_i = \sqrt{\frac{4}{3}\left|\frac{a_1}{a_3}\right|} \tag{9.91}$$

9.3.4 Blocker or Jammer

In a typical front-end amplifier shown in Fig. 9.19a, the low noise amplifier (LNA) is preceded by a channel filter to remove any unwanted interfering signals. The remaining spectrum at the input, though, may contain in-band blockers (other users of the channel). Consider the input spectrum of a weak desired signal and a “blocker” as shown in Fig. 9.19b

$$S_i = \underbrace{S_1\cos\omega_1 t}_{\text{Blocker}} + \underbrace{s_2\cos\omega_2 t}_{\text{Desired}} \tag{9.92}$$

We shall show that in the presence of a strong interferer, the gain of the system for the desired signal is reduced. This is true even if the interference signal is at a substantially difference frequency. We call this interference signal a “jammer”.

Obviously, the linear terms do not create any kind of desensitization. The second order terms, likewise, generate second harmonic and intermodulation, but not any fundamental signals. Te cubic term $a_3S_i^3$, though, generates the jammer desensitization term

$$S_i^3 = S_1^3\cos^3\omega_1 t + s_2^3\cos^3\omega_2 t + 3S_1^2 s_2\cos^2\omega_1 t\cos\omega_2 t + 3s_2^2 S_2\cos^2\omega_2 t\cos\omega_1 t \tag{9.93}$$

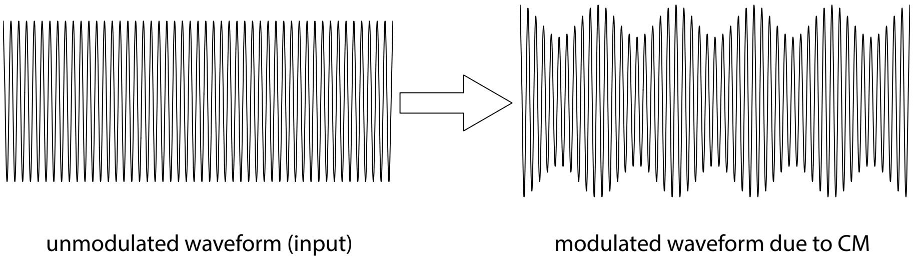


Figure 9.20: An unmodulated signal passed through a non-linear system picks-up modulation from another signal through cross-modulation.

The first two terms generate cubic and third harmonic whereas the last two terms generate fundamental signals at ω_1 and ω_2. The last term is much smaller as $s_2 \ll S_1$. The blocker term is therefore given by

$$a_3 3 S_1^2 s_2 \frac{1}{2} \cos \omega_2 t \tag{9.94}$$

This term adds or *subtracts* from the desired signal. The gain is therefore given by

$$\text{Apparent Gain} = \frac{a_1 s_2 + a_3 \frac{3}{2} S_1^2 s_2}{s_2} \tag{9.95}$$

Suppose that $a_3/a_1 < 0$, then we have

$$= a_1 + a_3 \frac{3}{2} S_1^2 = \underbrace{a_1}_{\text{small-signal gain}} \underbrace{\left(1 - \frac{3}{2}\left|\frac{a_3}{a_1}\right| S_1^2\right)}_{\text{blocker induced gain reduction}} \tag{9.96}$$

This result is very important as it shows that as long as the jammer signal passes through the amplifier, it reduces the gain for the desired signal, even if they are not in close frequency proximity.

Out of Band 3-dB Desensitization

Let's find the blocker power necessary to desensitize the amplifier by 3 dB. Solving the above equation

$$20 \log \left(1 - \frac{3}{2}\left|\frac{a_3}{a_1}\right| S_1^2\right) = -3\,\text{dB} \tag{9.97}$$

We find that the blocker power is given by

$$P_{OB} = P_{-1\,\text{dB}} + 1.2\,\text{dB} \tag{9.98}$$

It's now clear that we should avoid operating our amplifier with any signals in the vicinity of $P_{-1\,\text{dB}}$, since gain reduction occurs if the signals are any larger. At this signal level, there is also significant intermodulation distortion since $IP_3 = P_{-1\,\text{dB}} + 9.6\,\text{dB}$, which implies that $IM_3 \sim 2 \times 9.6\text{dBc} = 19.2\text{dBc}$.

9.3.5 Cross-Modulation of AM Signals

Consider a simple AM signal (modulated by a single tone)

$$s(t) = S_2(1 + m \cos \omega_m t) \cos \omega_2 t \tag{9.99}$$

where the modulation index $m \leq 1$. This can be written as

$$s(t) = S_2 \cos \omega_2 t + \frac{m}{2} \cos(\omega_2 - \omega_m)t + \frac{m}{2} \cos(\omega_2 + \omega_m)t \tag{9.100}$$

The first term is the RF carrier and the last terms are the modulation sidebands. Cross modulation (CM) occurs in AM systems (e.g. video cable tuners) when the modulation of a large AM signal transfers to another carrier going through the same amplifier, as shown in Fig. 9.20

$$S_i = \underbrace{S_1 \cos \omega_1 t}_{\text{wanted}} + \underbrace{S_2(1 + m \cos \omega_m t) \cos \omega_2 t}_{\text{interferer}} \tag{9.101}$$

CM occurs when the output contains a term like

$$K(1 + \delta \cos \omega_m t) \cos \omega_1 t \tag{9.102}$$

where δ is called the transferred modulation index. For $S_o = a_1 S_i + a_2 S_i^2 + a_3 S_i^3 + \cdots$, the term $a_2 S_i^2$ does not produce any CM. The term

$$a_3 S_i^3 = \cdots + 3a_3 S_1 \cos \omega_1 t \left(S_2(1 + m \cos \omega_m t) \cos \omega_2 t\right)^2 \tag{9.103}$$

is expanded to

$$= \cdots + 3a_3 S_1 S_2^2 \cos \omega_1 t(1 + 2m \cos \omega_m t + m^2 \cos^2 \omega_m t) \times \tfrac{1}{2}(1 + \cos 2\omega_2 t) \tag{9.104}$$

We can now identify the important term

$$S_o = \cdots + a_1 S_1 (1 + 3\frac{a_3}{a_1} S_2^2 m \cos \omega_m t) \cos \omega_1 t \tag{9.105}$$

We define the cross modulation as

$$CM = \frac{\text{Transferred Modulation Index}}{\text{Incoming Modulation Index}} \tag{9.106}$$

$$CM = 3\frac{a_3}{a_1} S_2^2 = 4IM_3 \tag{9.107}$$

We see that for a memoryless system, there is a relation bewteen CM, IM_3, and HD_3

$$= IM_3(\text{dB}) + 12\text{dB} \tag{9.108}$$

$$= 12HD_3 = HD_3(\text{dB}) + 22\text{dB} \tag{9.109}$$

9.3.6 Power Series Inversion

Sometimes it's easier to find a power series relation for the input in terms of the output. This happens when there is an analytical relation between the input and output that cannot be easily inverted. If we express this relation as a power series

$$S_i = a_1 S_o + a_2 S_o^2 + a_3 S_o^3 + \cdots \tag{9.110}$$

We can calculate the inverse relation

$$S_o = b_1 S_i + b_2 S_i^2 + b_3 S_i^3 + \cdots \tag{9.111}$$

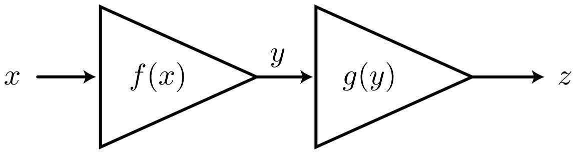


Figure 9.21: The cascade of two memoryless non-linear systems.

by substituting the above equation into the original equation and equating coefficient of like powers

$$S_i = a_1(b_1 S_i + b_2 S_i^2 + b_3 S_i^3 + \cdots) + a_2(\cdots)^2 + a_3(\cdots)^3 + \cdots \tag{9.112}$$

Equating linear terms, we find, as expected, that $a_1 b_1 = 1$, or $b_1 = 1/a_1$. Equating the square terms, we have

$$0 = a_1 b_2 + a_2 b_1^2 \tag{9.113}$$

$$b_2 = -\frac{a_2 b_1^2}{a_1} = -\frac{a_2}{a_1^3} \tag{9.114}$$

Finally, equating the cubic terms we have

$$0 = a_1 b_3 + a_2 2 b_1 b_2 + a_3 b_1^3 \tag{9.115}$$

$$b_3 = \frac{2a_2^2}{a_1^5} - \frac{a_3}{a_1^4} \tag{9.116}$$

It's interesting to note that if a power series does not have cubic, $a_3 \equiv 0$, the inverse series will nevertheless display cubic due to the first term above.

9.3.7 Amplifier Cascade

A common situation is that we cascade two non-linear systems, as shown in Fig. 9.21. For instance, a two-stage amplifier can be decomposed into the cascade such that

$$y = f(x) = a_1 x + a_2 x^2 + a_3 x^3 + \cdots \tag{9.117}$$

$$z = g(y) = b_1 y + b_2 y^2 + b_3 y^3 + \cdots \tag{9.118}$$

We'd like to find the overall relation

$$z = c_1 x + c_2 x^2 + c_3 x^3 + \cdots \tag{9.119}$$

To find c_1, c_2, $\cdots$, we simply substitute one power series into the other and collect like powers. The linear terms, as expected, are given by

$$c_1 = b_1 a_1 = a_1 b_1 \tag{9.120}$$

The square terms are given by

$$c_2 = b_1 a_2 + b_2 a_1^2 \tag{9.121}$$

The first term is simply the second order distortion produced by the first amplifier and amplified by the second amplifier linear term. The second term is the generation of second order by the second amplifier. Finally, the cubic terms are given by

$$c_3 = b_1 a_3 + b_2 2 a_1 a_2 + b_3 a_1^3 \tag{9.122}$$

The first and last term have a very clear origin. The middle terms, though, are more interesting. They arise due to second harmonic interaction when the second order distortion of the first amplifier can interact with the linear term through the second order non-linearity to produce cubic distortion. Even if both amplifiers have negligible cubic, $a_3 = b_3 \equiv 0$, we see the overall amplifier can generate cubic through this mechanism.

For example, consider two square law devices in cascade, such that $a_3 = b_3 = 0$ but both have a_2 and b_2. Now uppose that the first amplifier generates distortion products like $\omega_1 \pm \omega_2$ through second order non-linearity a_2. Now the second amplifier also will effectively multiply its own inputs through b_2, which multiplies this distortion product $\omega_1 \pm \omega_2$ with the fundamentals such as ω_1 and ω_2. This generates $2\omega_1 \pm \omega_2$ and $2\omega_2 \pm \omega_1$, 3rd order intermodulation products.

Example 22: Cascade Example

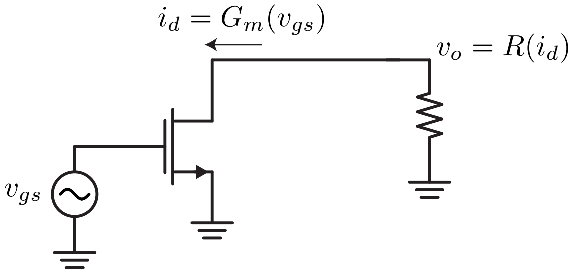


Figure 9.22: A MOS amplifier drives a non-linear load. This problem can be analyzed as a cascade of two non-linearities.

In most transistor amplifiers, we can decompose the non-linearity as a cascade of two non-linearities, as shown in Fig. 9.22. The G_m non-linearity

$$i_d = G_{m1} v_{in} + G_{m2} v_{in}^2 + G_{m3} v_{in}^3 + \cdots \tag{9.123}$$

and the output impedance non-linearity

$$v_o = R_1 i_d + R_2 i_d^2 + R_3 i_d^3 + \cdots \tag{9.124}$$

The output impedance can be a non-linear resistor load (such as a current source load) or simply the load of the device itself, which has a non-linear component.

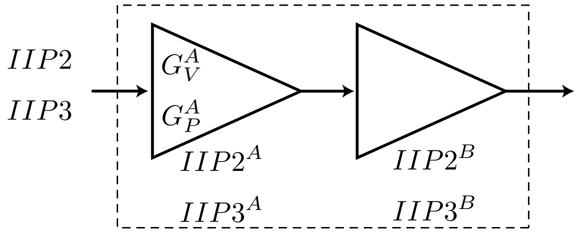


Figure 9.23: The setup for the calculation of the cascade intercept points.

$IIP2$ Cascade

Commonly we'd like to know the performance of a cascade in terms of the individual amplifier $IIP2$'s and $IIP3$'s, as shown in Fig. 9.23. To do this, note that $IIP2 = c_1/c_2$, so for the cascade we have (using our derived results)

$$\frac{c_2}{c_1} = \frac{b_1 a_2 + b_2 a_1^2}{b_1 a_1} = \frac{a_2}{a_1} + \frac{b_2}{b_1} a_1 \tag{9.125}$$

This leads to

$$\frac{1}{IIP2} = \frac{1}{IIP2^A} + \frac{a_1}{IIP2^B} \tag{9.126}$$

This is a very intuitive result, since it simply says that we can *input refer* the $IIP2$ of the second amplifier to the input by the voltage gain of the first amplifier.

In general, since the signal amplitude grows through a cascade of non-linearities, the distortion products generated downstream are increasingly important and tend to dominate the overall linearity of the system.

Example 23: $IIP2$ Cascade Example

Suppose the input amplifiers of a cascade has $IIP2^A = +0\,\text{dBm}$ and a voltage gain of 20 dB. The second amplifier has $IIP2^B = +10\,\text{dBm}$.

The input referred $IIP2_i^B = 10\,\text{dBm} - 20\,\text{dB} = -10\,\text{dBm}$. This is a much smaller signal than the $IIP2^A$, so clearly the second amplifier dominates the distortion. The overall distortion is given by $IIP2 \approx -12\,\text{dB}$.

Now suppose $IIP2^B = +20\,\text{dBm}$. Since $IIP2_i^B = 20\,\text{dBm} - 20\,\text{dB} = 0\,\text{dBm}$, we cannot assume that either amplifier dominates. Using the formula, we see the actual $IIP2$ of the cascade is degraded by a factor of 2, $IIP2 = -3\,\text{dBm}$.

$IIP3$ Cascade

Using the same approach, let's start with

$$\frac{c_3}{c_1} = \frac{b_1 a_3 + b_2 a_1 a_2 2 + b_3 a_1^3}{b_a a_1} = \left(\frac{a_3}{a_1} + \frac{b_3}{b_1} a_1^2 + \frac{b_2}{b_1} 2 a_2\right) \tag{9.127}$$

The last term, the second harmonic interaction term, will be neglected for simplicity. Then we have

$$\frac{1}{IIP3^2} = \frac{1}{IIP3_A^2} + \frac{a_1^2}{IIP3_B^2} \tag{9.128}$$

which shows that the $IIP3$ of the second amplifier is input referred by the voltage gain squared, or the power gain.

The second harmonic interaction term can be neglected when the two amplifiers are narrowband. In such a case, the second order non-linearity of the first amplifier is filtered and does not interact in a substantial way with the second amplifier. For example, if the first amplifier generates second harmonic, the narrowband filter attenuates these distortion products, preventing them from mixing with the fundamental to generate IM_3. Likewise, if the first amplifier generates IM_2 products, the narrowband filter prevents these distortion products from entering the second amplifier, where they can interact with the fundamental to produce mixing products. In this case, AC coupling also helps to attenuate these low frequency distortion products.

Example 24: LNA/Mixer Example

A common situation is an LNA and mixer cascade. The mixer can be characterized as a non-linear block with a given $IIP2$ and $IIP3$. Suppose the LNA has an $IIP3^A = -10\,\text{dBm}$ and a power gain of 20 dB and the mixer has an $IIP3^B = -20\,\text{dBm}$. If we input refer the mixer $IP3$, we have $IIP3_i^B = -20\,\text{dBm} - 20\,\text{dB} = -40\,\text{dBm}$. In this case, the mixer will dominate the overall $IIP3$ of the system.

9.3.8 *N* Tone Excitation of a MemoryLess Non-Linearity

Consider the effect of an m'th order non-linearity on an input of N tones

$$y_m = \left(\sum_{n=1}^{N} A_n \cos \omega_n t \right)^m \tag{9.129}$$

Let's use Eulor's identity again

$$y_m = \left(\sum_{n=1}^{N} \frac{A_n}{2} \left(e^{\omega_n t} + e^{-\omega_n t} \right) \right)^m \tag{9.130}$$

The above sum is over positive and negative frequencies. We can double the range of the sum by defining a negative frequency index as a negative frequency

$$y_m = \left(\sum_{n=-N}^{N} \frac{A_n}{2} e^{\omega_n t} \right)^m \tag{9.131}$$

where we assumed that $A_0 \equiv 0$ and $\omega_{-k} = -\omega_k$. The product of sums can be written as lots of sums

$$= \underbrace{\sum() \times \sum() \times \sum() \cdots \times \sum()}_{m-\text{times}} \tag{9.132}$$

$$= \sum_{k_1=-N}^{N} \sum_{k_2=-N}^{N} \cdots \sum_{k_m=-N}^{N} \frac{A_{k_1} A_{k_2} \cdots A_{k_m}}{2^m} \times e^{j(\omega_{k_1}+\omega_{k_2}+\cdots+\omega_{k_m})t} \tag{9.133}$$

Notice that we generate frequency component $\omega_{k_1} + \omega_{k_2} + \cdots + \omega_{k_m}$, sums and differences between m non-distinct frequencies. There are a total of $(2N)^m$ terms.

Let's take a simple example of $m = 3$, $N = 2$. We already know that this cubic non-linearity will generate harmonic distortion and IM products. We have $(2N)^m = 4^3 = 64$ combinations of complex frequencies, $\omega \in \{-\omega_2, -\omega_1, \omega_1, \omega_2\}$. There are 64 terms that looks like this (HD_3). For example, third harmonic is generated by

$$\omega_1 + \omega_1 + \omega_1 = 3\omega_1 \tag{9.134}$$

and IM_3 products are generated as follows

$$\omega_1 + \omega_1 + \omega_2 = 2\omega_1 + \omega_2 \tag{9.135}$$

and also

$$\omega_1 + \omega_1 - \omega_2 = 2\omega_1 - \omega_2 \tag{9.136}$$

Note that we can also generate fundamental, which corresponds to gain compression or expansion

$$\omega_1 + \omega_1 - \omega_1 = \omega_1 \tag{9.137}$$

Notice that from all of these 64 terms, many generate the same frequency so we now need to figure out how many are unique.

Frequency Mix Vector

Let the vector $\vec{k} = (k_{-N}, \cdots, k_{-1}, k_1, \cdots, k_N)$ be a $2N$-vector where element k_j denotes the number of times a particular frequency appears in a given term. As an example, consider the frequency terms

$$\left.\begin{array}{l} +\omega_2 - \omega_1 + \omega_2 = 2\omega_2 - \omega_1 \\ -\omega_1 + \omega_2 + \omega_2 = 2\omega_2 - \omega_1 \\ +\omega_2 + \omega_2 - \omega_1 = 2\omega_2 - \omega_1 \end{array}\right\} \vec{k} = (0,1,0,2) \tag{9.138}$$

all described by the same vector $\vec{k}$. Note that since $-\omega_1$ appears only once, there's a 1 in its entry. Likewise, since ω_2 appears twice, there's a 2 in its entry.

Properties of $\vec{k}$

First note that from Eq. (9.133), the sum of the k_j must equal m

$$\sum_{j=-N}^{N} k_j = k_{-N} + \cdots + k_{-1} + k_1 + \cdots + k_N = m \tag{9.139}$$

where m is the power of the non-linearity.

For a fixed vector $\vec{k_0}$, how many different sum vectors are there? We can sum m frequencies $m!$ ways, but since the order of the sum is irrelevant, we need to divide out the suplicities. Since each k_j coefficient can be ordered $k_j!$ ways, the number of ways to form a given frequency product is given by the number of times this occurs

$$(m;\vec{k}) = \frac{m!}{(k_{-N})! \cdots (k_{-1})!(k_1)! \cdots (k_N)!} \tag{9.140}$$

Since our signal is real, each term has a complex conjugate present. Hence there is another vector $\vec{k}'_0$ given by

$$\vec{k}'_0 = (k_N, \cdots, k_1, k_{-1}, \cdots, k_{-N}) \tag{9.141}$$

Notice that the components are in reverse order since $\omega_{-j} = -\omega_j$. If we take the sum of these two terms we have

$$2\Re\left\{e^{j(\omega_{k_1}+\omega_{k_2}+\cdots+\omega_{k_m})t}\right\} = 2\cos(\omega_{k_1}+\omega_{k_2}+\cdots+\omega_{k_m})t \tag{9.142}$$

The amplitude of a frequency product is thus given by 2 times the number of times it occurs (Eq. 9.140) scaled by the factor $1/2^m$, or

$$\frac{2\times(m;\vec{k})}{2^m} = \frac{(m;\vec{k})}{2^{m-1}} \tag{9.143}$$

Example 25:

***IM*₃ Revisited**

Using this new tool, let's derive an equation for the IM_3 product more directly. Since we have two tones, $N = 2$. IM_3 is generated by the $m = 3$ non-linear term. A particular IM_3 product, such as $(2\omega_1 - \omega_2)$, is generated by the frequency mix vector $\vec{k} = (1,0,2,0)$.

$$(m;\vec{k}) = \frac{3!}{1!\cdot 2!} = 3 \tag{9.144}$$

$$2^{m-1} = 2^2 = 4 \tag{9.145}$$

So the amplitude of the IM_3 product is $3/4a_3s_i^3$. Relative to the fundamental

$$IM_3 = \frac{3}{4}\frac{a_3s_i^3}{a_1s_i} = \frac{3}{4}\frac{a_3}{a_1}s_i^2 \tag{9.146}$$

Example 26:

Pentic Non-Linearity

Let's calculate the gain expansion/compression due to the 5th order non-linearity. For a one tone, we have $N = 1$ and $m = 5$. A pentic term generates fundamental as follows

$$\omega_1 + \omega_1 + \omega_1 - \omega_1 - \omega_1 = \omega_1 \tag{9.147}$$

In terms of the $\vec{k}$ vector, this is captured by $\vec{k} = (2,3)$. The amplitude of this term is given by

$$(m;\vec{k}) = \frac{5!}{2!\cdot 3!} = \frac{5\cdot 4}{2} = 10 \tag{9.148}$$

$$2^{m-1} = 2^4 = 16 \tag{9.149}$$

So the fundamental amplitude generated is $a_5 \frac{10}{16} S_i^5$.

Example 27:

Apparent Gain Due to Pentic

The apparent gain of the system, including the 3rd and 5th, is thus given by

$$\text{AppGain} = a_1 + \frac{3}{4} a_3 S_i^2 + \frac{10}{16} a_5 S_i^4 \tag{9.150}$$

At what signal level is the 5th order term as large as the 3rd order term?

$$\frac{3}{4} a_3 S_i^2 = \frac{10}{16} a_5 S_i^4 \tag{9.151}$$

$$S_i = \sqrt{1.2 \frac{a_3}{a_5}} \tag{9.152}$$

For a bipolar amplifier, we found that $a_3 = 1/(3!V_t^3)$ and $a_5 = 1/(5!V_t^5)$. Solving for S_i, we have

$$S_i = V_t \sqrt{1.2 \times 5 \times 4} \approx 127\,\text{mV} \tag{9.153}$$

9.4 Effect of Feedback on Distortion

The block diagram of a feedback amplifier is shown in Fig. 16.11. In this section we will show that in most situations, the action of the feedback linearizes a system, and thus reduces the distortion generated by a non-linear device. We know this to be true for large loop gain, since in the limit of infinite loop gain, the closed-loop gain of such a system only depends on the feedback network, which is usually realized with linear passive elements rather than active devices.

Assume that the only distortion is in the forward path a

$$s_o = a_1 s_\varepsilon + a_2 s_\varepsilon^2 + a_3 s_\varepsilon^3 + \cdots \tag{9.154}$$

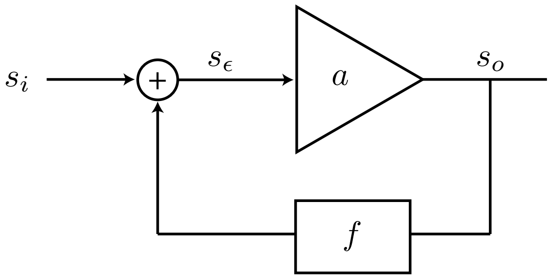


Figure 9.24: A generic feedback amplifier.

where s_ε is the difference signal applied to the non-linear device

$$s_\varepsilon = s_i - f s_o \tag{9.155}$$

where $f < 1$ is the feedback factor. Substituting the above expression results in Eq. 9.154

$$s_o = a_1(s_i - f s_o) + a_2(s_i - f s_o)^2 + a_3(s_i - f s_o)^3 + \cdots \tag{9.156}$$

Ultimately we'd like to derive an equation in the following form

$$s_o = b_1 s_i + b_2 s_i^2 + b_3 s_i^3 + \cdots \tag{9.157}$$

Substitute this desired solution into the equation to obtain

$$\begin{aligned} b_1 s_i &+ b_2 s_i^2 + b_3 s_i^3 + \cdots = a_1(s_i - f b_1 s_i - f b_2 s_i^2 - \cdots) \\ &+ a_2(b_1 s_i + b_2 s_i^2 + b_3 s_i^3 + \cdots)^2 + a_3(b_1 s_i + b_2 s_i^2 + b_3 s_i^3 + \cdots)^3 + \cdots \end{aligned} \tag{9.158}$$

Solve for the first order terms

$$b_1 s_i = a_1(s_i - f b_1 s_i) \tag{9.159}$$

$$b_1 = \frac{a_1}{1 + a_1 f} = \frac{a_1}{1 + T} \tag{9.160}$$

The above equation is the same as linear analysis (loop gain $T = a_1 f$). Now let's equate second order terms

$$b_2 s_i^2 = -a_1 f b_2 s_i^2 + a_2(s_i - f b_1 s_i)^2 \tag{9.161}$$

$$b_2(a + a_1 f) = a_2 \left(1 - \frac{f a_1}{1 + T}\right)^2 \tag{9.162}$$

$$b_2(1 + T)^3 = a_2(1 + T - T)^2 = a_2 \tag{9.163}$$

$$b_2 = \frac{a_2}{(1 + T)^3} \tag{9.164}$$

Equating third-order terms, we arrive at

$$b_3 s_i^3 = a_1(-fb_3 s_i^3) + a_2(-fb_2 2s_i^3) + a_3(s_i - fb_1 s_i)^3 + \cdots \tag{9.165}$$

$$b_3(1 + a_1 f) = -2a_2 b_2 f \frac{1}{1+T} + \frac{a_3}{(1+T)^3} \tag{9.166}$$

$$b_3(1+T) = \frac{-2a_2 f}{1+T}\frac{a_2}{(1+T)^3} + \frac{a_3}{(1+T)^3} \tag{9.167}$$

$$b_3 = \frac{a_3(1+a_1 f) - 2a_2^2 f}{(1+a_1 f)^5} \tag{9.168}$$

Second Order Interaction

The term $2a_2^2 f$ is the second order interaction. Second order distortion in the forward path is fed back and combined with the input linear terms to generate third order distortion. We can eliminate the third order distortion if

$$a_3(1+a_1 f) = 2a_2^2 f \tag{9.169}$$

Harmonic Distortion in Feedback Amplifiers

The harmonic distortion in a feedback amplifier is now easily calculated using the modified coefficients of the power series

$$HD_2 = \frac{1}{2}\frac{b_2}{b_1^2} s_{om} \tag{9.170}$$

$$= \frac{1}{2}\frac{a_2}{(1+T)^3}\frac{(1+T)^2}{a_1^2} s_{om} \tag{9.171}$$

$$= \frac{1}{2}\frac{a_2}{a_1^2}\frac{s_{om}}{1+T} \tag{9.172}$$

Recall that without feedback $HD_2 = \frac{1}{2}\frac{a_2}{a_1^2} s_{om}$. For a *given output* signal, the negative feedback reduces the second order distortion by $\frac{1}{1+T}$.

Likewise the third harmonic distortion ratio can be calculated

$$HD_3 = \frac{1}{4}\frac{b_3}{b_1^3} s_{om}^2 \tag{9.173}$$

$$= \frac{1}{4}\frac{a_3(1+T) - 2a_2^2 f}{(1+T)^5}\frac{(1+T)^3}{a_1^3} s_{om}^2 \tag{9.174}$$

$$= \underbrace{\frac{1}{4}\frac{a_3}{a_1^3} s_{om}^2}_{\text{disto with no fb}} \frac{1}{1+T}\left[1 - \frac{2a_2^2 f}{a_3(1+T)}\right] \tag{9.175}$$

Note that in most of the equations derived so far, the improvement in linearity comes with loop gain T. It turns out that the same equations hold true at high frequency if we replace the DC value of T with $T(j\omega)$.

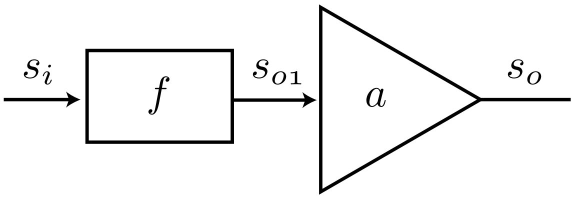


Figure 9.25: A non-linear block preceded by a linear attenuator.

Feedback versus Input Attenuation

Notice that the distortion is improved for a given output signal level, which is what we want. We can simply improve the input referred distortion by using an attenuator as shown in Fig. 9.25. Say $s_{o1} = f s_i$ with $f < 1$, then

$$s_o = a_1 s_{o1} + a_2 s_{o1}^2 + a_3 s_{o1}^3 + \cdots = \underbrace{a_1 f}_{b_1} s_i + \underbrace{a_2 f^2}_{b_2} s_i^2 + \underbrace{a_3 f^3}_{b_3} s_i^3 + \cdots \tag{9.176}$$

Note that all the non-linear coefficents are reduced and even b_2/b_1 and b_3/b_1 are lower, so for the same input signal level, there's less distortion. There's a price to pay in that the noise performanc eof the amplifier has been comprised (see Sec. **??** in Ch. **??**). But the distortion is unchanged for a given output signal

$$HD_2 = \frac{1}{2}\frac{b_2}{b_1^2} s_{om} = \frac{1}{2}\frac{a_2}{a_1^2} s_{om} \tag{9.177}$$

In other words, if we want to drive the same power into a load, the distortion is the same since we have to raise the input of the amplifier back up to the same level. With feedback, though, we actually improve the distortion, even for the same output level.

9.4.1 Bipolar Amplifiers With Feedback

In this section we will apply the derivations of the previous section to practical amplifiers. It is not always obvious how to map a real amplifier into the ideal model represented by Fig. 16.11. In some case, such a simple representation may not be valid at all. But fortunately most feedback amplifiers can be approximated in such a manner.

Bipolar Amplifier with Emitter Degeneration

Consider a common example, a common emitter amplifier with degeneration shown in Fig. 9.26. Previously we calculated the distortion without feedback and found the amplifier to be extremely non-linear due to the exponential current response. The action of the feedback is to reduce the non-linearity by sensing the output current and feeding back an input voltage which subtracts from the v_{be} of the device.

Working directly with the equations, the total input signal applied to the base of the amplifier is

$$v_i + V_Q = V_{BE} + I_E R_E \tag{9.178}$$

The V_{BE} and I_E terms can be split into DC and AC currents (assume $\alpha \approx 1$)

$$v_i + V_Q = V_{BE,Q} + v_{be} + (I_Q + i_c) R_E \tag{9.179}$$

Subtracting bias terms we have a separate AC and DC equation

$$V_Q = V_{BE,Q} + I_Q R_E \tag{9.180}$$

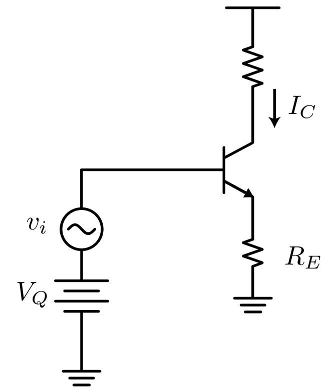


Figure 9.26: A common emitter amplifier with degeneration.

$$v_i = v_{be} + i_C R_E \tag{9.181}$$

The AC equation can be put into the following form

$$v_{be} = v_i - i_c R_E \tag{9.182}$$

Compare this to our feedback equation

$$s_\varepsilon = s_i - f s_o \tag{9.183}$$

The equations have the same form with the following substitutions

$$s_\varepsilon = v_{be} \tag{9.184}$$

$$s_o = i_c \tag{9.185}$$

$$s_i = v_i \tag{9.186}$$

$$f = R_E \tag{9.187}$$

Now we know that

$$i_c = a_1 v_{be} + a_2 v_{be}^2 + a_3 v_{be}^3 + \cdots \tag{9.188}$$

where the coefficients $a_{1,2,3,\cdots}$ come from expanding the exponential into a Taylor series

$$a_1 = g_m \;\; a_2 = \frac{1}{2}\frac{I_Q}{V_t^2} \;\; \cdots \tag{9.189}$$

With feedback we have

$$i_c = b_1 v_i + b_2 v_i^2 + b_3 v_i^3 + \cdots \tag{9.190}$$

The loop gain $T = a_1 f = g_m R_E$, which allows us to write

$$b_1 = \frac{g_m}{1+g_m R_E} \tag{9.191}$$

$$b_2 = \frac{\frac{1}{2}\left(\frac{q}{kT}\right)^2 I_Q}{(1+g_m R_E)^3} \tag{9.192}$$

$$b_3 = \cdots \tag{9.193}$$

Harmonic Distortion with Feedback

Using our derived modified coefficients we have

$$HD_2 = \frac{1}{2}\frac{b_2}{b_1^2} s_{om} \tag{9.194}$$

$$= \frac{1}{4}\frac{\hat{i_c}}{i_q}\frac{1}{1+g_m R_E} \tag{9.195}$$

$$HD_3 = \frac{1}{4}\frac{b_3}{b_1^3} s_{om}^2 \tag{9.196}$$

$$= \frac{1}{24}\left(\frac{\hat{i_c}}{I_Q}\right)^2 \frac{1-\frac{3g_m R_E}{1+g_m R_E}}{1+g_m R_E} \tag{9.197}$$

Harmonic Distortion Null

You may notice something interesting in Eq. 9.197. Since the second term in the numerator is strictly positive (passive resistor), we can adjust the feedback to obtain a null in HD_3. Solve $HD_3 = 0$

$$\frac{3g_m R_E}{1+g_m R_E} = 1 \tag{9.198}$$

which gives us the critical value of R_E

$$R_E = \frac{1}{2g_m} \tag{9.199}$$

For example, for $I_Q = 1\text{mA}$, $R_E = 13\Omega$. This is plotted in Fig. 9.27.

Bipolar Amplifier with Finite Source Resistance

Up to now we have assumed a perfect “voltage" driven amplifier. In reality, all sources have non-zero source impedance which we must take into account. Re-writing the KVL equation for the input loop (Fig. 9.28)

$$v_i + V_Q - I_B R_B = V_{BE} + I_E R_E \tag{9.200}$$

Assume that $\alpha \approx 1$, $\beta = \beta_0$ (constant). Let $R_B = R_S + r_b$ represent the total resistance at the base

$$v_i + V_Q = V_{BE} + I_C\left(R_E + \frac{R_B}{\beta_0}\right) \tag{9.201}$$

The formula is the same as the case of a BJT with emitter degeneration with $R'_E = R_E + R_B/\beta_0$. In other words, the source resistance also tends to linearize the BJT amplifier, but with a feedback factor that is β times smaller, since the base current is directly proportional to the collector (output) current.

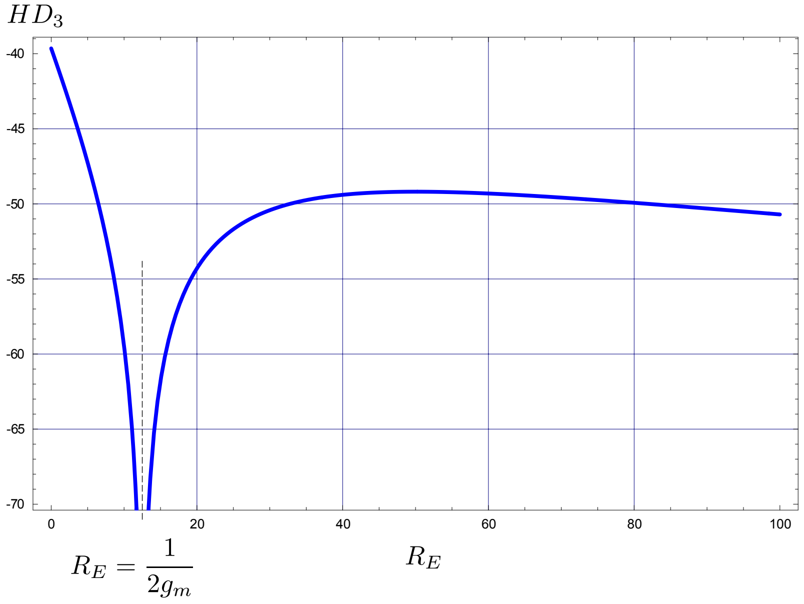


Figure 9.27: The third-order harmonic distortion as a function of R_E for a common emitter amplifier.

Emitter Follower

For an emitter follower shown in Fig. 9.29a, we notice that the same equations apply as before with $R_E = R_L$. If the load resistance is large enough, then the loop gain is very large and the emitter follower is very linear. We see that the amount of distortion generated by the follower is a strong function of the load resistance and the required drive.

Common Base Amplifier

Finally, for the common base amplifier shown in Fig. 9.29b, we also note that the same equation applies with R_E acting as the feedback element

$$v_i - V_Q + I_C R_E = -V_{BE} \tag{9.202}$$

Therefore the same equations can be used. In many cases R_E is the source resistance and so the action of the source linearizes the amplifier. This amplifier is more linear than a common emitter amplifier driven with the same source resistance since the current through R_E is roughly the same as the collector (output) current.

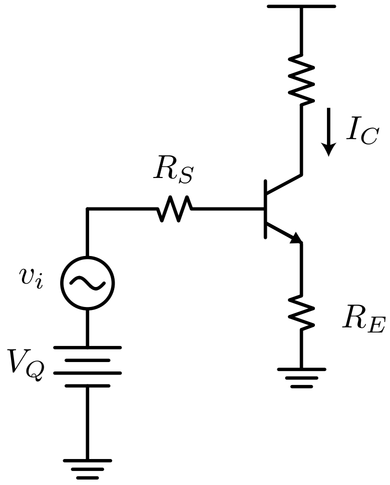


Figure 9.28: A common emitter amplifier driven with source resistance.

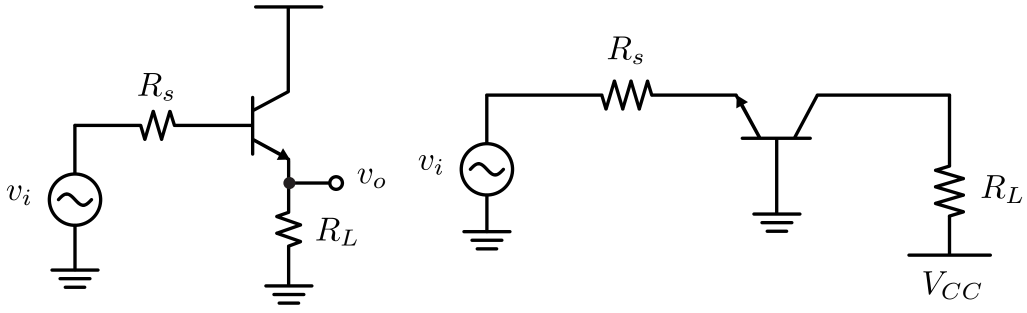
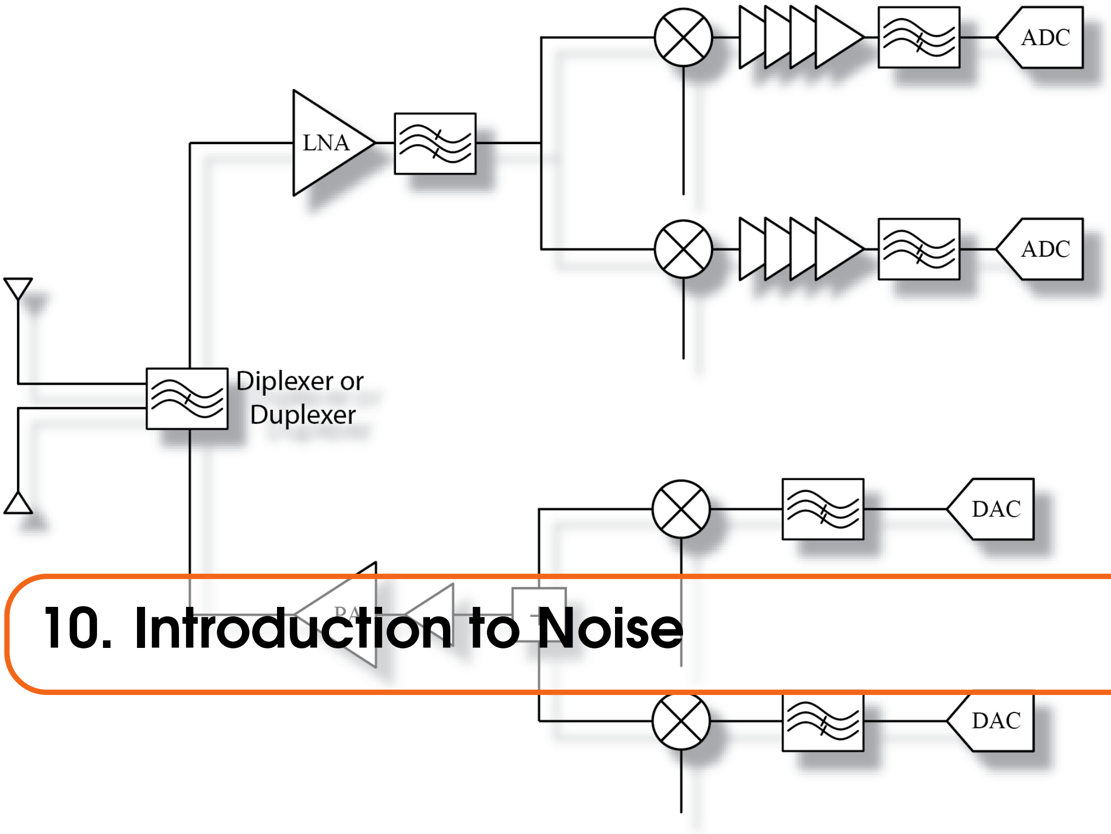


Figure 9.29: Setup for the calculation of the distortion in a (a) emitter follower amplifier and (b) common base amplifier.



10. Introduction to Noise

As shown in Fig. 10.1, even in the absence of an input signal, all electronic amplifiers generate noise. This noise originates from the random thermal motion of carriers and the discreteness of charge. Noise signals are random and must be treated by statistical means. Even though we cannot predict the actual noise waveform, we can predict the statistics such as the mean (average) and power in a specific frequency band of interest.

10.0.1 Noise Power

The average value of the noise waveform is zero

$$\overline{v_n}(t) =< v_n(t) >= \frac{1}{T}\int_T v_n(t)dt = 0 \tag{10.1}$$

The mean is also zero, which means if we freeze time and take an infinite number of samples from identical amplifiers, the average voltage from all samples will be zero. The variance, though, is non-zero. Equivalently, we may say that the signal power is non-zero

$$\overline{v_n(t)^2} = \frac{1}{T}\int_T v_n^2(t)dt \neq 0 \tag{10.2}$$

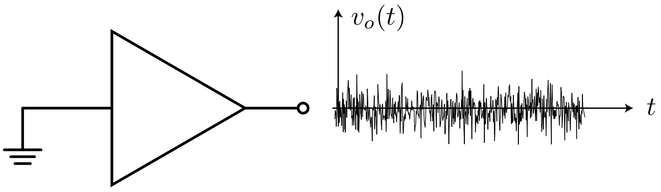


Figure 10.1: A hypothetical amplifier with grounded input generates noise at the output due to internal noise sources.

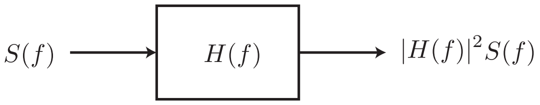


Figure 10.2: Say something.

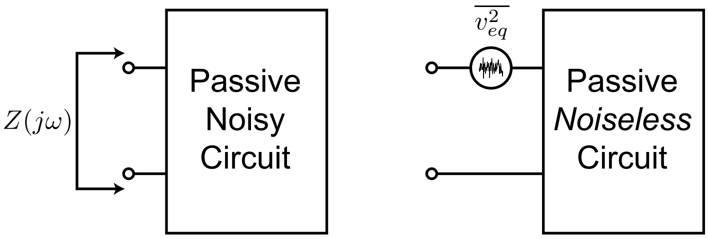


Figure 10.3: A noisy circuit can be replaced by a noiseless circuit and an equivalent noise generator.

The RMS (root-mean-square) voltage is given by

$$v_{n,rms} = \sqrt{\overline{v_n(t)^2}} \tag{10.3}$$

10.1 Noise in an LTI System

A fundamental result from Stochastic Systems, derived in the following section, is that if you inject noise into an LTI system (such as a filter), the output noise is shaped by the magnitude of the transfer function

$$\overline{V}^2 = \int_{-\infty}^{\infty} S(f)|H(f)|^2 df \tag{10.4}$$

Note that we can't say anything about the phase, but we know the magnitude response will be filtered. Any white noise source, such as a resistor, will be shaped by poles in the system.

10.1.1 Noise for Passive Circuits

For a general linear circuit, such as the N port shown in Fig. 11.4, the mean square noise voltage (current) at any port is given by the equivalent input resistance (conductance)

$$\overline{v_{eq}^2} = 4kT\Re(Z(f))\delta f \tag{10.5}$$

This is the "spot" noise. If the network has a filtering property, then we integrate over the band of interest

$$\overline{v_{T,eq}^2} = 4kT\int_B \Re(Z(f))df \tag{10.6}$$

Unlike resistors, L's and C's *do not* generate noise. They do shape the noise due to their frequency dependence.

10.1.2 Noise Analysis

To find the equivalent mean square noise voltage for a circuit, we use the small signal-model (noise signals are actually small, so it's a great approximation). For each noise source, invoke

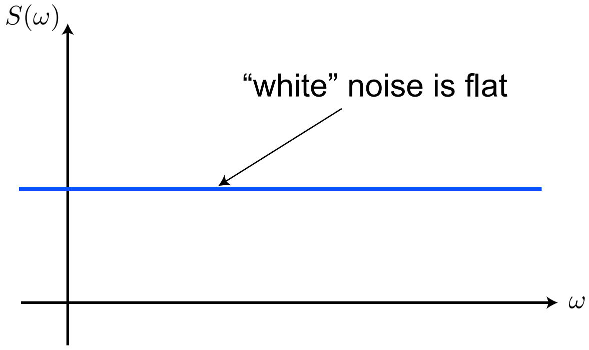


Figure 10.4: The power spectrum of a white noise source.

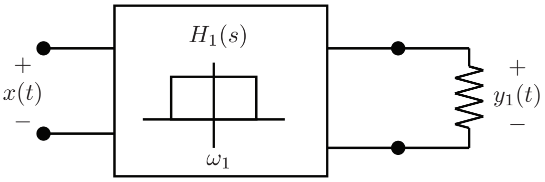


Figure 10.5: Measurement setup to measure the noise power in a particular frequency band.

superposition and calculate the noise contribution to the desired node

$$\overline{v_{n,o}^2} = |G_{1,o}|^2\overline{v_{n,1}^2} + |G_{2,o}|^2\overline{v_{n,2}^2} + \cdots = \sum_k |G_{k,o}|^2\overline{v_{n,k}^2} \tag{10.7}$$

where $\overline{v_{n,k}^2}$ is the kth noise source, and the gain from that noise to the output node is given by $G_{k,o}$. Note that the polarity of the noise sources is irrelevant since we're summing powers (all positive quantities). The above expression assumes that the noise sources are *independent*. Later on we'll see how to handle correlated noise sources.

10.1.3 Power Spectrum of Noise (*Optional*)

The power spectrum of the noise shows the concentration of noise power as a function of frequency. Many noise sources are "white" in that the spectrum is flat (up to extremely high frequencies), as shown in Fig. 10.4. In such cases the noise waveform is totally unpredictable as a function of time. In other words, there is absolutely no correlation between the noise waveform at time t_1 and some later time $t_1 + \delta$, no matter how small we make δ. To understand the power spectrum of a signal, let us motivate it through a simple experiment. Suppose we have a power meter, a device that measures the power of a signal by computing a time average of the square of a signal

$$p = \int_T x(t)^2 dt \tag{10.8}$$

To measure the amount of power at a particular frequency band, then, we would use a sharp filter around the frequency of interest and pass the signal into the pwoer meter, as shown in Fig. 10.5. To see the corresponding mathematical operation, let us calculate the average power of a random signal that is applied to a filter with impulse reponse $h_1(t)$

$$\overline{y_1{}^2(t)} = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\left(\int_{-\infty}^{\infty} h_1(\tau_1)x(t-\tau_1)\,d\tau_1\right)\left(\int_{-\infty}^{\infty} h_1(\tau_2)x(t-\tau_2)\,d\tau_2\right)dt \tag{10.9}$$

Re-arranging the order of integration

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_1(\tau_1)\, h_1(\tau_2) \left(\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} x(t-\tau_1)\, x(t-\tau_2)\, dt \right) d\tau_1 d\tau_2 \tag{10.10}$$

The autocorrelation function is defined as

$$\varphi_{xx}(t) = \overline{x(t)\, x(t+\tau)} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)\, x(t+\tau)\, dt \tag{10.11}$$

which allows us to re-write the above average power expression as

$$\overline{{y_1}^2(t)} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_1(\tau_1)\, h_2(\tau_2) \varphi_{xx}(\tau_1-\tau_2)\, d\tau_1 d\tau_2 \tag{10.12}$$

Consider the Fourier Transform pair

$$\varphi_{xx}(j\omega) = \int_{-\infty}^{\infty} \varphi_{xx}(\tau)\, e^{-j\omega\tau} d\tau \tag{10.13}$$

and

$$\varphi_{xx}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_{xx}(j\omega)\, e^{j\omega\tau} d\omega \tag{10.14}$$

$\varphi_{xx}(j\omega)$ is a real and even function of ω since $\varphi_{xx}(t)$ is a real and even function of t. Substitute the frequency domain representation of φ into the expression for the average power

$$\overline{{y_1}^2(t)} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_1(\tau_1)\, h_1(\tau_2) \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_{xx}(j\omega)\, e^{j\omega(\tau_1-\tau_2)} d\omega d\tau_1 d\tau_2 \tag{10.15}$$

Again re-arrange the order of integration

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_{xx}(j\omega) \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_1(\tau_1)\, h_1(\tau_2)\, e^{j\omega(\tau_1-\tau_2)} d\tau_1 d\tau_2 \tag{10.16}$$

Note that the integrations over τ are separable into two Fourier transforms

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_{xx}(j\omega) \left(\int_{-\infty}^{\infty} h_1(\tau_1)\, e^{+j\omega\tau_1} d\tau_1 \right) \left(\int_{-\infty}^{\infty} h_1(\tau_2)\, e^{-j\omega\tau_2} d\tau_2 \right) d\omega \tag{10.17}$$

or

$$\overline{{y_1}^2(t)} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_{xx}(j\omega)\, H_1(j\omega)\, {H_1}^*(j\omega)\, d\omega \tag{10.18}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_{xx}(j\omega)\, |H_1(j\omega)|^2 d\omega \tag{10.19}$$

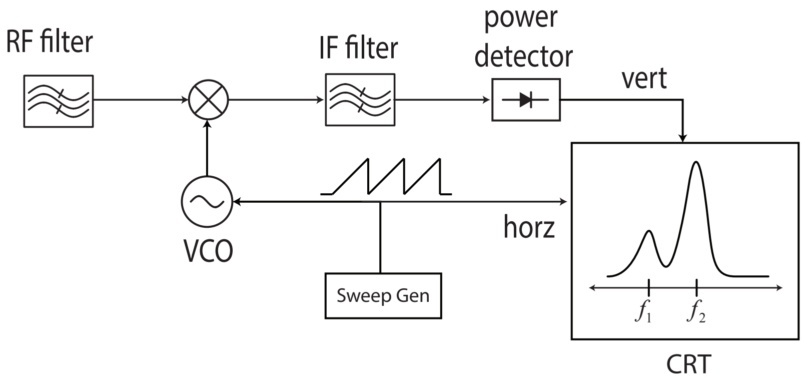


Figure 10.6: Two equivalent representations of the noise of a resistor.

The average power in the frequency range $0 < f < f_1$ is now

$$\overline{y_1{}^2(t)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_{xx}(j\omega)|H_1(j\omega)|^2 d\omega \tag{10.20}$$

For a brick-wall filter

$$= \frac{1}{2\pi}\int_{-\omega_1}^{\omega_1} \varphi_{xx}(j\omega)\,d\omega = \int_{-f_1}^{f_1} \varphi_{xx}(j2\pi f)\,df = 2\int_{0}^{f_1} \varphi_{xx}(j2\pi f)\,df \tag{10.21}$$

We can now generalize and say that to measure the power in any frequency range, apply an ideal bandpass filter with passband $f_1 < f < f_2$

$$\overline{y_1{}^2(t)} = 2\int_{f_1}^{f_2} \varphi_{xx}(j2\pi f)\,df \tag{10.22}$$

The interpretation of φ_{xx} as the the *power spectral density* (PSD) is clear.

10.1.4 Spectrum Analyzer

A spectrum analyzer measures the PSD of a signal. We can build a “Poor Man’s" spectrum analyzer as shown in Fig. 10.6 by utilizing a fixed IF or baseband high Q filter and a mixer to down-convert the band of interest to the IF. By sweeping the frequency of the mixer LO port, we can sweep the frequency that is down-converted, and thus by plotting the output voltage over time, we are viewing the power spectrum of the signal. The plot can be generated with a CRT, with the vertical controlled by the power detector and the horizontal controlled by the sweep ramp waveform.

A modern spectrum analyzer is similar in spirit but uses an analog-to-digital converter instead of a power detector so that the bandwidth of the PSD observation can be changed on the fly.

10.2 Thermal Noise

10.2.1 Resistor Thermal Noise

All resistors generate noise. The noise power generated by a resistor R can be represented by a series voltage source with mean square value $\overline{v_n^2}$

$$\overline{v_n^2} = 4kTRB \tag{10.23}$$

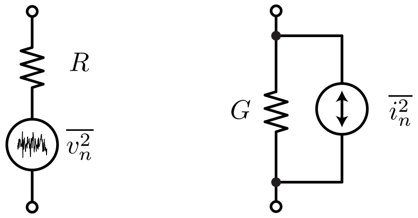


Figure 10.7: Two equivalent representations of the noise of a resistor.

Equivalently, we can represent this with a current source in shunt

$$\overline{i_n^2} = 4kTGB \tag{10.24}$$

Here B is the bandwidth of observation and kT is Boltzmann's constant times the temperature of observation. An equivalent circuit is shown in Fig. 10.7, where the noisy resistor is replaced by an ideal noiseless resistor and a noise generator. This result comes from thermodynamic considerations, thus explaining the appearance of kT. Often we speak of the "spot noise", or the noise in a specific narrowband δf

$$\overline{v_n^2} = 4kTR\delta f \tag{10.25}$$

Since the noise is white, the shape of the noise spectrum is determined by the external elements (L's and C's).

Example 28: Resistor Noise Example

Suppose that $R = 10\text{k}\Omega$ and $T = 20°\text{C} = 293\text{K}$.

$$4kT = 1.62 \times 10^{-20} \tag{10.26}$$

$$\overline{v_n^2} = 1.62 \times 10^{-16} \times B \tag{10.27}$$

$$v_{n,rms} = \sqrt{\overline{v_n(t)^2}} = 1.27 \times 10^{-8}\sqrt{B} \tag{10.28}$$

If we limit the bandwidth of observation to $B = 10^6$MHz, then we have

$$v_{n,rms} \approx 13\mu\text{V} \tag{10.29}$$

This represents the limit for the smallest voltage we can resolve across this resistor in this bandwidth

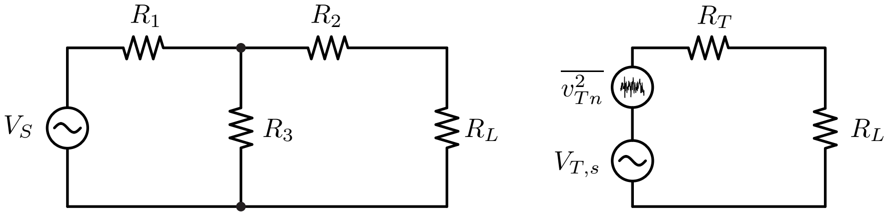


Figure 10.8: A complicated resistive circuit can be simplified using equivalent generators that have noise.

Combination of Resistors

If we put two resistors in series, then the mean square noise voltage is given by

$$\overline{v_n^2} = 4kT(R_1 + R_2)B = \overline{v_{n1}^2} + \overline{v_{n2}^2} \tag{10.30}$$

The noise powers add, *not* the noise voltages. Likewise, for two resistors in parallel, we can add the mean square currents

$$\overline{i_n^2} = 4kT(G_1 + G_2)B = \overline{i_{n1}^2} + \overline{i_{n2}^2} \tag{10.31}$$

This holds for any pair of independent noise sources (zero correlation).

For an arbitrary resistive circuit shown in Fig. 10.8, we can find the equivalent noise by using a Thevenin (Norton) equivalent circuit or by transforming all noise sources to the output by the appropriate *power* gain (e.g. voltage squared or current squared)

$$V_{T,s} = V_S \frac{R_3}{R_1 + R_3} \tag{10.32}$$

$$\overline{v_{Tn}^2} = 4kTR_TB = 4kT(R_2 + R_1||R_3)B \tag{10.33}$$

Example 29: Noise of an RC Circuit

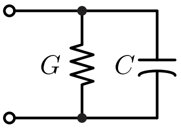


Figure 10.9: A simple *RC* circuit generates a total integrated noise independent of the resistance.

To find the equivalent mean square noise voltage of an *RC* circuit shown in Fig. 10.9, begin by calculating the impedance

$$Z = \frac{1}{Y} = \frac{1}{G + j\omega C} = \frac{G - j\omega C}{G^2 + \omega^2 C^2} \tag{10.34}$$

Integrating the noise over all frequencies, we have

$$\overline{v_n^2} = \frac{4kT}{2\pi} \int_0^\infty \frac{G}{G^2 + \omega^2 C^2} d\omega = \frac{kT}{C} \tag{10.35}$$

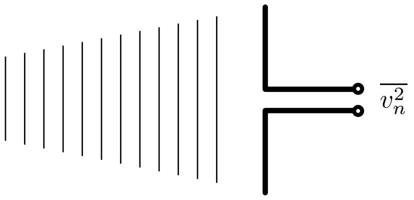


Figure 10.10: The noise power radiation incident on an ideal antenna produces an equivalent noise voltage at the terminals.

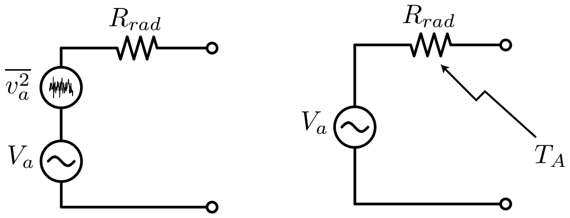


Figure 10.11: The equivalent circuit of an antenna includes the radiation resistance and the associated noise.

> Notice the result is *independent* of R. Since the noise and bandwidth are proportional/inversely proportional to R, the influence of R cancels out.

10.2.2 Noise of a Receiving Antenna

Assume we construct an antenna with ideal conductors so $R_{wire} = 0$ (Fig. 10.10). If we connect the antenna to a spectrum analyzer, though, we will observe noise. The noise is also “white” but the magnitude depends on where we point our antenna (sky versus ground).

$$\overline{v_a^2} = 4kT_A R_{rad} B \tag{10.36}$$

T_A is the equivalent antenna temperature and R_{rad} is the radiation resistance of the antenna. Since the antenna does not generate any of its own thermal noise, the observed noise must be incident on the antenna. In fact, it’s “black body” radiation. Physically T_A is related to the temperature of the external bodies radiating into space (e.g. space or the ground). We can represent the noise of the antenna with the equivalent circuit shown in Fig. 10.11.

10.3 Physical Origin of Noise (*Optional*)

An elegant derivation of the physical origin of the noise of a resistor is due to van der Ziel. Consider an RC circuit where as a result of thermal agitation of electrons, the capacitor is charged and discharged constantly. On average, the energy stored is given by the equipartition theorem:

$$\frac{1}{2}C\overline{V^2} = \frac{1}{2}k_B T \tag{10.37}$$

$$\overline{V^2} = \frac{k_B T}{C} \tag{10.38}$$

This result will be derived rigorously assuming a Boltzmann distribution for the energy.

10.3.1 Noise Voltage Due to Resistor

Let's say that we don't know the power spectral density of the noise of the resistor. Whatever it's noise is, though, we know that noise voltage at the capacitor can be computed from

$$\overline{V}^2 = \int_{-\infty}^{\infty} S_V^2(\omega)|H(\omega)|^2 d\omega \tag{10.39}$$

where $H(\omega)$ is the transfer function from the resistor noise to the capacitor

$$H(\omega) = \frac{1}{1 + j\omega RC} \tag{10.40}$$

$$|H(\omega)|^2 = \frac{1}{1 + \omega^2(RC)^2} \tag{10.41}$$

Integrating the noise we have

$$\overline{V}^2 = \int_{-\infty}^{\infty} S_V^2(\omega) \frac{1}{1 + \omega^2(RC)^2} d\omega \tag{10.42}$$

If we assume that the voltage noise density does not depend on frequency (experimental fact), then we have

$$\overline{V}^2 = \overline{S}_V^2 \int_{-\infty}^{\infty} \frac{1}{1 + \omega^2(RC)^2} d\omega = \frac{S_V^2}{2RC} \tag{10.43}$$

Now applying the Equipartition Theorem

$$\frac{\overline{S_V}^2}{2RC} = \frac{k_B T}{C} \tag{10.44}$$

$$\overline{S_V}^2 = 2k_B TR \tag{10.45}$$

In most noise calculations, we integrate noise over positive frequencies, which means we should double the result of our previous calculation to properly account for noise

$$\overline{S_V}^2 = 4k_B TR \tag{10.46}$$

This is the result that we quoted earlier, first observed by Johnson and derived analytically by Nyquist using a different Thermodynamic argument.

Energy Stored in a Capacitor Due to Noise

We will now prove the earlier result where we invoked the equipartition theorem. For a capacitor, the energy stored $E = CV^2/2$ due to a noise resistor should be proportional to the Boltzmann distribution $\exp(-E/k_BT)$ (assume thermal equilibrium at temperature T). To find the proportionality constant, note that integrating this quantity over all energy values should be unity

$$\int_{-\infty}^{\infty} K \exp\left(\frac{-CV^2}{2k_B T}\right) dV = 1 \tag{10.47}$$

so

$$K = \sqrt{\frac{C}{2\pi k_B T}} \tag{10.48}$$

Now we can compute the mean squared value of the voltage

$$\overline{V^2} = \sqrt{\frac{C}{2\pi k_B T}} \int_{-\infty}^{\infty} V^2 \exp\left(\frac{-CV^2}{2k_B T}\right) dV \tag{10.49}$$

$$\overline{V^2} = \frac{k_B T}{C} \tag{10.50}$$

The last step follows after performing the integral.

10.4 Noise in Active Devices

10.4.1 Diode Shot Noise

A forward biased diode exhibits noise called *shot noise*. This noise arises due to the quantized nature of charge crossing a potential barrier. The noise mean square current is given by

$$\overline{i_{d,n}^2} = 2qI_{DC}B \tag{10.51}$$

The noise is white and proportional to the DC current I_{DC}. Reversed biased diodes exhibit excess noise not related to shot noise.

10.4.2 Noise in a Bipolar Junction Transistor

All physical resistors in a BJT produce noise (r_b, r_e, r_c). The output resistance r_o, though, is *not* a physical resistor. Likewise, r_π, is not a physical resistor. Thus these resistances do not generate noise since they are modeling elements rather than actual physical resistors. Similar to a diode, the junctions of a BJT exhibit shot noise

$$\overline{i_{b,n}^2} = 2qI_B B \tag{10.52}$$

$$\overline{i_{c,n}^2} = 2qI_C B \tag{10.53}$$

At low frequencies the transistor exhibits “Flicker Noise” or $1/f$ Noise.

BJT Hybrid-Π Model

The extended equivalent circuit shown in Fig. 10.12 includes noise sources. Note that a small-signal equivalent circuit is appropriate because the noise perturbation is very small and in most cases the non-linear effects can be ignored. On the other, the time-varying effects of the circuit play a very important role in communication circuits, a topic we will visit when we study mixers.

10.4.3 FET Noise

In addition to the extrinsic physical resistances in a FET (r_g, r_s, r_d), the channel resistance also contributes thermal noise. The drain current noise of the FET is therefore given by

$$\overline{i_{d,n}^2} = 4kT\gamma g_{ds0}\delta f + K\frac{I_D^a}{C_{ox}L_{eff}^2 f^e}\delta f \tag{10.54}$$

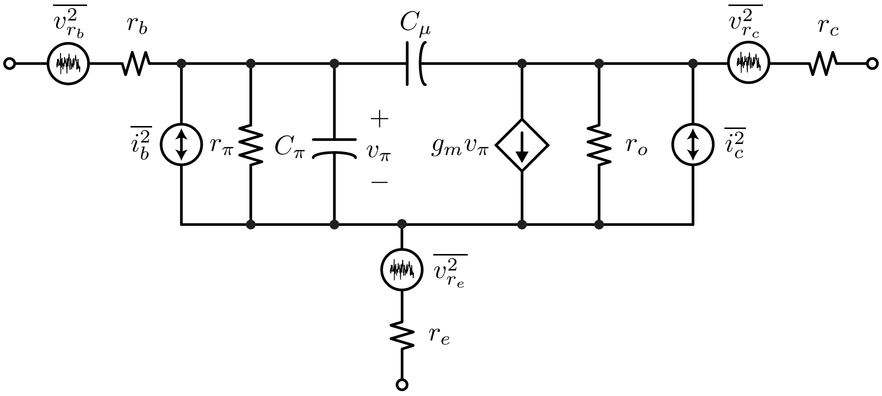


Figure 10.12: The equivalent circuit of a bipolar amplifier with noise sources shown explicitly.

The first term is the thermal noise due to the channel resistance and the second term is the “Flicker Noise”, also called the $1/f$ noise, which dominates at low frequencies. The factor $\gamma = \frac{2}{3}$ for a long channel device. The constants K, a, and e are usually determined empirically. FETs typically exhibit much higher flicker noise, an artifact of the way these devices are fabricated. In a FET where the current conduction occurs at the surface, so called “dangling bonds" and oxide traps contribute to the flicker noise.

FET Channel Resistance

Consider a FET with $V_{DS} = 0$. Then the channel conductance is given by

$$g_{ds,0} = \frac{\partial I_{DS}}{\partial V_{DS}} = \mu C_{ox}\frac{W}{L}(V_{GS} - V_T) \tag{10.55}$$

For a long-channel device, this is also equal to the device transconductance g_m in saturation

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = \mu C_{ox}\frac{W}{L}(V_{GS} - V_T) \tag{10.56}$$

For short-channel devices, this relation is not true, but we can define

$$\alpha = \frac{g_m}{g_{d0}} \neq 1 \tag{10.57}$$

FET Noise Equivalent Circuit

The equivalent circuit shown in Fig. 10.13 includes the noise sources discussed thus far. Additionally, the resistance of the substrate also generates thermal noise. In most circuits we will be concerned with the noise due to the channel $\overline{i_d^2}$ and the input gate noise $\overline{v_{R_g}^2}$ dominate.

10.5 Noise in Communication Systems

10.5.1 Degradation of Link Quality

As we have seen, noise is an ever present part of all systems. Since transmitters and receivers are made with electronic components, they add even more noise to the signals that we are trying to transmit and detect. In analog systems, noise deteriorates the quality of the received signal, e.g. the appearance of “snow” on the TV screen, or “static” sounds during an audio transmission. In digital communication systems, noise can occasionally introduce errors in the transmitter bits, and so it degrades the Bit Error Rate (BER). This ultimately results in a reduced throughput because it requires retransmission of data packets or extra coding to recover the data in the presence of errors.

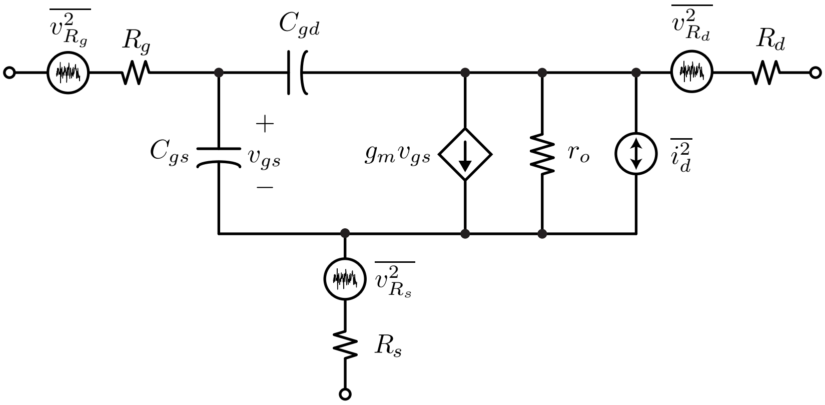


Figure 10.13: The equivalent circuit of a FET with the noise sources shown explicitly.

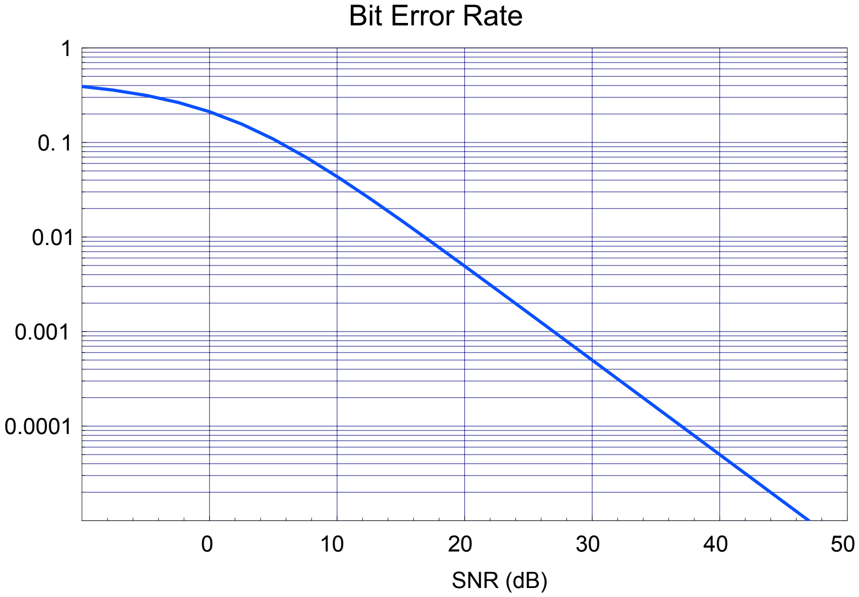


Figure 10.14: The Bit Error Rate (*BER*) versus the signal-to-noise ratio (*SNR*) for a hypotheical communication system.

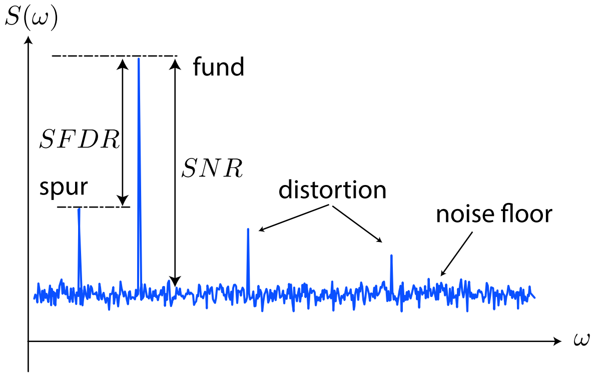


Figure 10.15: The spurious free dynamic range $SFDR$ measures the available dynamic range of a signal at a particular point in a system. For instance, in an amplifier the largest signal determines the distortion "noise" floor and the noise properties of the amplifier determine the "noise floor".

It's typical to plot the Bit-Error-Rate (BER) in a digital communication system for a given modulation scheme (Fig. 10.14). This shows the average rate of errors for a given signal-to-noise-ratio (SNR). For instance a $BER = 10^{-6}$ means that only 1 in a million bits will be detected in error. The more complicated the modulation scheme, the more SNR is required to faithfully transmit and receive these signals. Usually digital systems are quite tolerant of noise up a limit, which is shown as the "waterfall" point on the graph. For the modulation scheme shown in Fig. 10.14, the probably of error is nearly unity for an $SNR \sim 0\,\text{dB}$, which makes sense. On the other hand, if the system requires a raw $BER < 10^{-3}$, then the required $SNR > 30\,\text{dB}$. In other words, the signal should be a thousand times stronger than the noise power.

In general, therefore, we strive to maximize the signal to noise ratio in a communication system. If we receive a signal with average power P_{sig}, and the average noise power level is P_{noise}, then the SNR is simply

$$SNR = \frac{P_{sig}}{P_{noise}} \tag{10.58}$$

$$SNR(\text{dB}) = 10 \cdot \log \frac{P_{sig}}{P_{noise}} \tag{10.59}$$

Here we distinguish between random noise and "noise" due to interferers or distortion generated by the amplifier. But a more complete description requires us to consider both effects. We can get away by ignoring the distortion if we assume that the input signal is weak (close to the level of noise) so that the distortion products are small. But in a worst case scenario, the weak signal is accompanied by an large interfering signal, which also produces significant distortion products. This is captured by the spurious free dynamic range $SFDR$, shown in Fig. 10.15. Often we call tones at the harmonics or intermodulation frequencies as distortion products and any other tones as "spurs". Often spurs are also harmonics of other signals, such as a reference clock.

10.5.2 Noise Figure

The *Noise Figure* (NF) of an amplifier is a block (e.g. an amplifier) is a measure of the degradation of the SNR, or the ratio the input signal-to-noise (SNR_i) to output signal-to-noise (SNR_o)

$$F = \frac{SNR_i}{SNR_o} \tag{10.60}$$

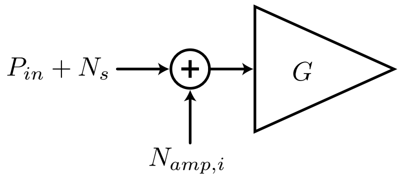


Figure 10.16: The noise of an amplifier can be equated to an equivalent input referred noise source and a noiseless amplifier. The input noise source is by definition a noise source that produces the same output noise as the actual amplifier.

Commonly quoted on a log scale

$$NF(\text{dB}) = 10 \cdot \log(F) \tag{10.61}$$

Since any system can at best pass the input noise to the output, $F \geq 1$ or $NF \geq 0\,\text{dB}$. But most systems add noise and thus the output noise is even more corrupt. The noise figure is measured (or calculated) by specifying a standard input noise level through the source resistance R_s at a given temperature. For RF communication systems, this is usually specified as $R_s = 50\Omega$ and $T = 293°K$.

Suppose an amplifier has a gain G and apply the definition of F

$$SNR_i = \frac{P_{sig}}{N_s} \tag{10.62}$$

$$SNR_o = \frac{GP_{sig}}{GN_s + N_{amp,o}} \tag{10.63}$$

The term $N_{amp,o}$ is the total output noise due to the amplifier in absence of any input noise.

$$SNR_o = \frac{P_{sig}}{N_s + \frac{N_{amp,o}}{G}} \tag{10.64}$$

Let $N_{amp,i}$ denote the total *input referred noise* of the amplifier (see Fig. 10.16)

$$SNR_o = \frac{P_{sig}}{N_s + N_{amp,i}} \tag{10.65}$$

The noise figure is therefore

$$F = \frac{SNR_i}{SNR_o} = \frac{\cancel{P_{sig}}}{N_s} \times \frac{N_s + N_{amp,i}}{\cancel{P_{sig}}} \tag{10.66}$$

$$F = 1 + \frac{N_{amp,i}}{N_s} \geq 1 \tag{10.67}$$

It's now even more clear that all amplifiers have a noise figure ≥ 1. Any real system degrades the *SNR* since all circuit blocks add noise.

The amount of noise added by the amplifier is normalized to the incoming noise N_s in the calculation of F. For RF systems, this is the noise of a 50Ω source at 293°K. Since any amplification degrades the *SNR*, why do any amplification at all? Because often the incoming signal is too weak to be detected without amplification.

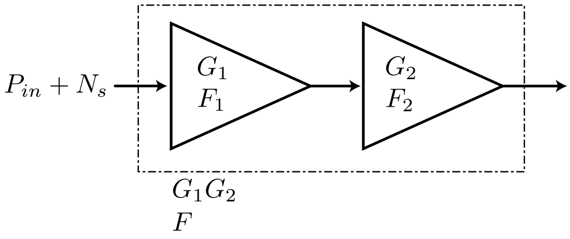


Figure 10.17: The noise figure of the cascade of two blocks can be readily calculated if the blocks are impedance matched to a common Z_0.

10.5.3 Noise Figure of Cascaded Blocks

If two blocks are cascaded as shown in Fig. 10.17, we would like to derive the noise figure of the total system. Assume the blocks are impedance matched properly to result in a gain $G = G_1G_2$. For each amplifier in cascade, we have

$$F_i = 1 + \frac{N_{amp,i}}{N_s} \tag{10.68}$$

By definition, the noise added by each amplifier to the input is given by

$$N_{amp,i} = N_s(F-1) \tag{10.69}$$

where N_s represents some standard input noise. If we now input refer all the noise in the system we have

$$N'_{amp,i} = N_s(F_1-1) + \frac{N_s(F_2-1)}{G_1} \tag{10.70}$$

which gives us the total noise figure of the amplifier

$$F = 1 + \frac{N'_{amp,i}}{N_s} = 1 + (F_1-1) + \frac{F_2-1}{G_1} = F_1 + \frac{F_2-1}{G_1} \tag{10.71}$$

Apply the formula to the last two blocks

$$F_{23} = F_2 + \frac{F_3-1}{G_2} \tag{10.72}$$

$$F = F_1 + \frac{F_{23}-1}{G_1} \tag{10.73}$$

$$= F_1 + \frac{F_2-1}{G_1} + \frac{F_3-1}{G_1G_2} \tag{10.74}$$

The general equation is written by inspection

$$= F_1 + \frac{F_2-1}{G_1} + \frac{F_3-1}{G_1G_2} + \frac{F_4-1}{G_1G_2G_3} + \cdots \tag{10.75}$$

It's important to realize that we have assumed that the noise powers can be moved easily from block to block without consideration of any impedances. This is true if the blocks are all matched to a common impedance.

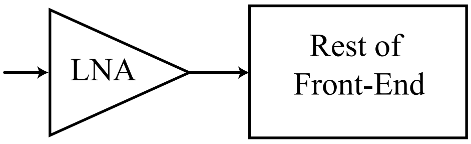


Figure 10.18: The low noise amplifier, or LNA, is the first block in any communication system. It is used to reject the noise of the rest of the system.

Low Noise Amplifiers

We see that in a cascade, the noise contribution of each successive stage is smaller and smaller. The noise of the *first* stage is the most important.Thus, every communication system employs a *low noise amplifier* (LNA) at the front to relax the noise requirements. A typical LNA might have a $G = 20\,\text{dB}$ of gain and a noise figure $NF < 1.5\,\text{dB}$. The noise figure depends on the application.

Example 30: NF Cascade Example

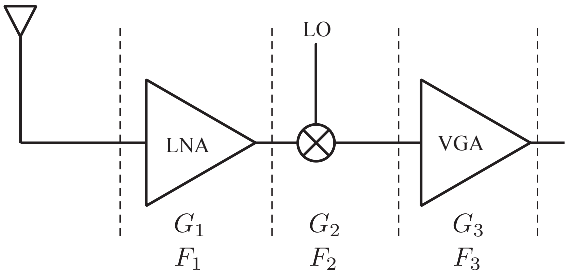


Figure 10.19: A typical receiver chain consists of an LNA, mixer, and IF amplifier.

Consider a typical receiver chain shown in Fig. 10.19, which includes an LNA with gain $G = 15\,\text{dB}$ and noise figure $NF = 1.5\,\text{dB}$, followed by a mixer with conversion gain of $G = 10\,\text{dB}$ and $NF = 10\,\text{dB}$. The IF amplifier has $G = 70\,\text{dB}$ and $NF = 20\,\text{dB}$. Even though the blocks operate at different frequencies, we can still apply the cascade formula if the blocks are impedance matched. The overall noise figure is given by

$$F = 1.413 + \frac{10-1}{60} + \frac{100-1}{60 \cdot 10} = 2.4\,\text{dB} \tag{10.76}$$

10.5.4 Minimum Detectable Signal

The minimum detectable signal (MDS) is the smallest signal that can appear at the input of the receiver and still result in a “proper" detection. Proper is a probabilistic concept and refers to a given desired value of the BER. Let’s do a simple example to make this concept more concrete. Suppose that a system requires an *SNR* of $10\,\text{dB}$ for proper detection (to satisfy the BER). Also suppose that the detector needs a minimum voltage amplitude of 10mV (for the analog-to-digital conversion). If a front-end receiver has a $NF = 10\,\text{dB}$, let's compute the minimum input power that can support communication:

$$SNR_o = \frac{SNR_i}{F} = \frac{\frac{P_{min}}{N_s}}{F} > 10 \tag{10.77}$$

or

$$P_{min} > 10 \cdot F \cdot N_s = 10 \cdot F \cdot kTB \tag{10.78}$$

we see that the answer depends on the bandwidth B.

$$P_{min} = 10\,\text{dB} + NF - 174\,\text{dBm} + 10 \cdot \log B \tag{10.79}$$

For wireless data, $B \sim 10\text{MHz}$:

$$P_{in} = 10\,\text{dB} + 10\,\text{dB} - 174\,\text{dB} + 70\,\text{dB} = -84\,\text{dBm} \tag{10.80}$$

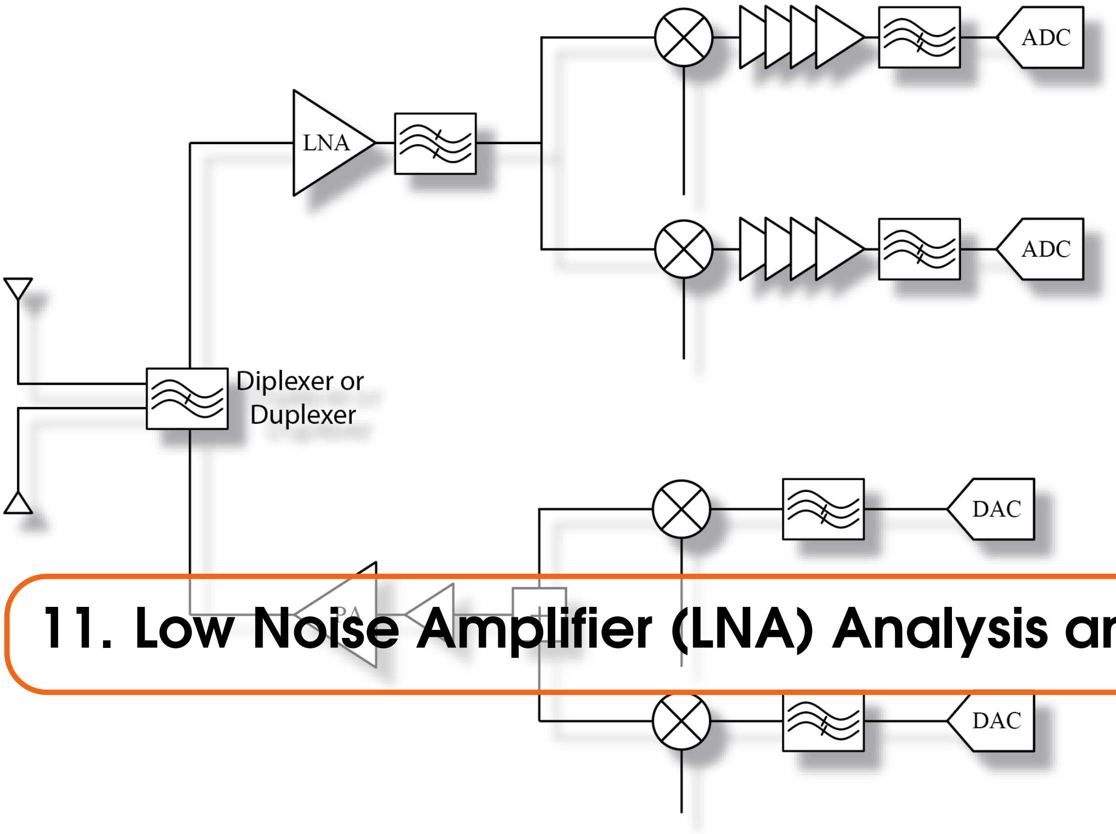
Translated into voltage on $50\,\Omega$, that's a signal of only $20\,\mu\text{V}$. Since the detector requires a signal of amplified $10\,\text{mV}$, the system should have a voltage gain of

$$G_V = \frac{10\,\text{mV}}{20\,\mu\text{V}} = 0.5 \times 10^3 \approx 500(54\,\text{dB}) \tag{10.81}$$

We see that the noise figure has a dB for dB impact on the minimum detectable input signal. Since the received power drops (at best) $> 20\,\text{dB}$ per decade of distance, a few dB improvement in NF may dramatically improve the coverage area of a communication link. Otherwise the transmitter has to boost the TX power, which requires excess power consumption due to the efficiency η of the transmitter.

10.6 Conclusion

We started this chapter by looking at the fundamental sources of noise and mathematical tools to analyze the noise power spectrum. Since noise is not a deterministic process, we need special tools to analyze noise problems. The physical origin of noise through thermal, shot, and flicker noise processes was discussed. Next we took a system perspective on noise and defined analyzed the noise figure of a communication system and the impact on the bit-error rate. Fundamentally, in an interference free environment, the sensitivity of a receiver is determined by its noise figure, which means it's very important to realize a front-end receiver with a low noise block that can interface with the antenna, providing sufficient gain with only a small degradation in the signal SNR. The design of such an low noise amplifier will be the topic of the next chapter.



11. Low Noise Amplifier (LNA) Analysis and Design

11.1 Two-Port Noise

Any noisy two port can be replaced with a *noiseless* two-port and equivalent input noise sources as shown in Fig. 11.1. In general, these noise sources are correlated. For now let's neglect the correlation. The equivalent sources are found by opening and shorting the input. As shown in Fig. 11.2a, when we short the input, the equivalent current noise does not enter the two-port whereas the voltage noise is applied across the input terminal of the two-port. By equating the output noise of the noisey and noiseless blocks, we can derive the magnitude of the input noise generator. Likewise, as shown in Fig. 11.2b, when we open circuit the two-port, the equivalent input noise generator is "dangling" and does not produce any noise whereas the full current noise flows into the two-port.

Example 31: BJT Noise Sources

Consider the equivalent circuit model of a common emitter BJT shown in Fig. 11.3.

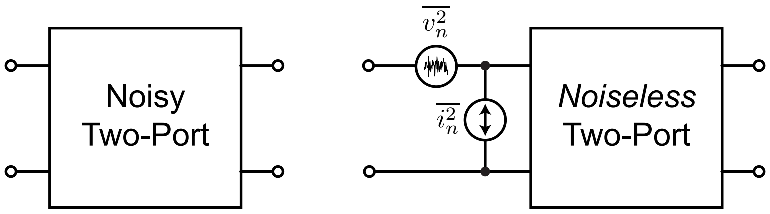


Figure 11.1: Any two port can be represented by an equivalent noiseless two-port and two correlated noise generators. A convenient representation involves the pair of input noise current/voltage generators.

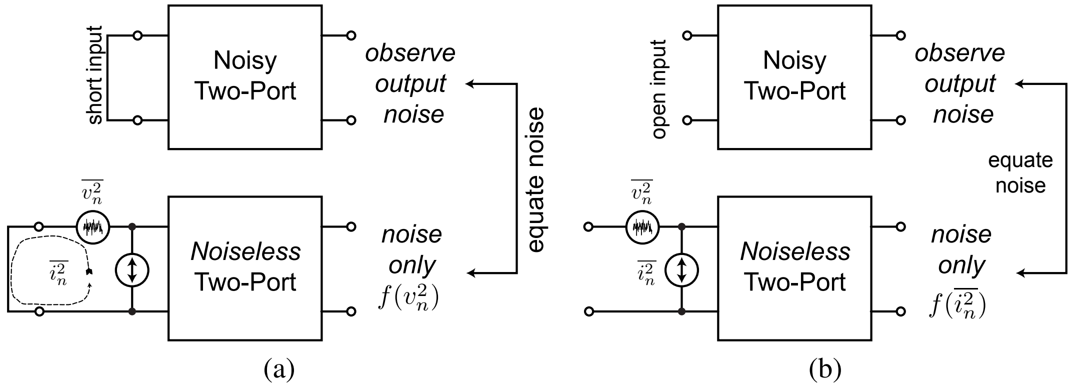


Figure 11.2: Procedure for calculating the equivalent input noise (a) voltage and (b) current.

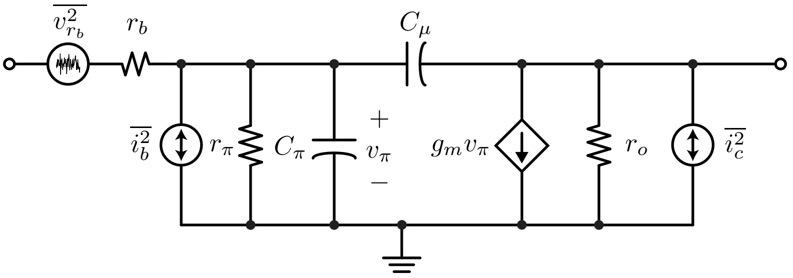


Figure 11.3: A simplified equivalent circuit model of the common emitter BJT including the noise sources.

Consider the base as port 1 and the collector as port 2 of a two-port circuit. Let us derive the equivalent generator noise voltage and current. If we leave the base of a BJT open, then the total output noise is given by

$$\overline{i_o^2} = \overline{i_c^2} + \beta^2 \overline{i_b^2} = \overline{i_n^2}\beta^2 \tag{11.1}$$

or

$$\overline{i_n^2} = \frac{\overline{i_c^2}}{\beta^2} + \overline{i_b^2} \approx \overline{i_b^2} \tag{11.2}$$

If we short the input of the BJT, we have

$$\overline{i_o^2} \approx g_m^2 \overline{v_n^2} \left(\frac{Z_\pi}{Z_\pi + r_b} \right)^2 = \beta^2 \frac{\overline{v_n^2}}{(Z_\pi + r_b)^2} \tag{11.3}$$

$$= \beta^2 \frac{\overline{v_{r_b}^2}}{(Z_\pi + r_b)^2} + \overline{i_c^2} \tag{11.4}$$

Solving for the equivalent BJT noise voltage

$$\overline{v_n^2} = \overline{v_{r_b}^2} + \frac{\overline{i_c^2}(Z_\pi + r_b)^2}{\beta^2} \tag{11.5}$$

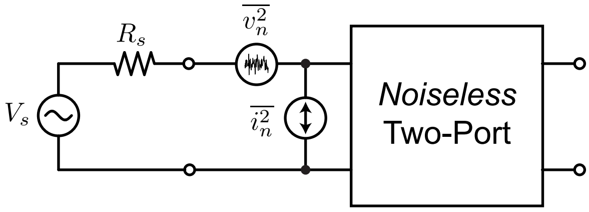


Figure 11.4: A two-port system represented by its equivalent input noise sources.

$$\overline{v_n^2} \approx \overline{v_{r_b}^2} + \frac{\overline{i_c^2} Z_\pi^2}{\beta^2} \tag{11.6}$$

At low frequencies this simplifies to

$$\overline{v_n^2} \approx \overline{v_{r_b}^2} + \frac{\overline{i_c^2}}{g_m^2} \tag{11.7}$$

$$\overline{v_n^2} = 4kTBr_b + \frac{2qI_CB}{g_m^2} \tag{11.8}$$

$$\overline{i_n^2} = \frac{2qI_c}{\beta} \tag{11.9}$$

11.1.1 Role of Source Resistance

We can now begin to appreciate that the generator source impedance plays an important role in the noise figure of a two port. Consider Fig. 11.4 where if we assume that $R_s = 0$, then only the voltage noise $\overline{v_n^2}$ is important. Likewise, if $R_s = \infty$, only the current noise $\overline{i_n^2}$ is important. Intuitively then, we can select an amplifier based on the following selection criteria: If R_s is large, then select an amplifier with low $\overline{i_i^2}$ (MOS input stage). If R_s is low, pick an amplifier with low $\overline{v_n^2}$ (BJT input stage). For a given R_s, there is an optimal $\overline{v_n^2}/\overline{i_n^2}$ ratio. Alternatively, for a given amplifier, there is an optimal R_s

Equivalent Input Noise Voltage

For a given application, the source impedance R_s is fixed. For many communication systems, the source impedance $R_s = 50\,\Omega$. In Fig. 11.4, let us find the total output noise voltage for a fixed R_s

$$\overline{v_o^2} = (\overline{v_n^2}A_v^2 + \overline{v_{R_s}^2}A_v^2)\left(\frac{R_{in}}{R_{in}+R_s}\right)^2 + \left(\frac{R_{in}}{R_{in}+R_s}\right)^2 R_s^2\overline{i_n^2}A_v^2 \tag{11.10}$$

$$= (\overline{v_n^2} + \overline{i_n^2}R_s^2 + \overline{v_{R_s}^2})\left(\frac{R_{in}}{R_{in}+R_s}\right)^2 A_v^2 \tag{11.11}$$

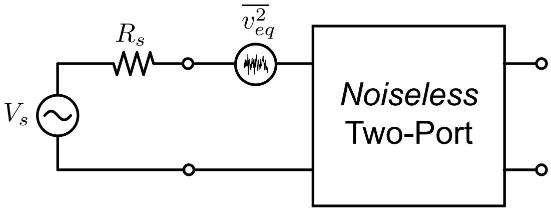


Figure 11.5: A general two-port can be simplified to a noise-less two-port and an equivalent voltage noise generator.

This means that we can simplify the equivalent circuit to Fig. 11.5, which captures the total noise into a simple equivalent voltage noise generator

$$\overline{v_{eq}^2} = \overline{v_n^2} + \overline{i_n^2} R_s^2 \tag{11.12}$$

Applying the definition of noise figure

$$F = 1 + \frac{N_{amp,i}}{N_s} = 1 + \frac{\overline{v_{eq}^2}}{\overline{v_s^2}} \tag{11.13}$$

Optimal Source Impedance

Let $\overline{v_n^2} = 4kTR_nB$ and $\overline{i_n^2} = 4kTG_nB$. Then

$$F = 1 + \frac{R_n + G_n R_s}{R_s} = 1 + G_n R_s + \frac{R_n}{R_s} \tag{11.14}$$

This equation is very revealing since it shows that there are only two noise terms in the above equation, one that scales with R_s and the other that scales inversely with R_s. Therefore there must be an optimum, which occurs when these terms contribute equally to the output noise. This is easy to prove by taking a derivative

$$\frac{dF}{dR_s} = G_n - \frac{R_n}{R_s^2} = 0 \tag{11.15}$$

We see that the noise figure is minimized for

$$R_{opt} = \sqrt{\frac{R_n}{G_n}} = \sqrt{\frac{\overline{v_n^2}}{\overline{i_n^2}}} \tag{11.16}$$

The major assumption we made was that $\overline{v_n^2}$ and $\overline{i_n^2}$ are not correlated. The resulting minimum noise figure is thus

$$F_{min} = 1 + G_n R_s + \frac{R_n}{R_s} \tag{11.17}$$

$$= 1 + G_n \sqrt{\frac{R_n}{G_n}} + \sqrt{\frac{G_n}{R_n}} R_n \tag{11.18}$$

$$= 1 + 2\sqrt{R_n G_n} \tag{11.19}$$

For an arbitrary amplifier the noise figure is F. Consider the difference between F and F_{min}

$$F - F_{min} = G_n R_s + \frac{R_n}{R_s} - 2\sqrt{R_n G_n} \tag{11.20}$$

$$= \frac{R_n}{R_s}(1 + \frac{G_n R_s^2}{R_n} - 2\frac{R_s}{R_n}\sqrt{R_n G_n} \tag{11.21}$$

$$= \frac{R_n}{R_s}\left(1 + \left(\frac{R_s}{R_{opt}}\right)^2 - \frac{2R_s}{R_{opt}}\right) \tag{11.22}$$

$$= \frac{R_n}{R_s}\left|\frac{R_s}{R_{opt}} - 1\right|^2 \tag{11.23}$$

$$= R_n R_s \left|G_{opt} - G_s\right|^2 \tag{11.24}$$

Sometimes R_n is called the *noise sensitivity parameter* since we can express the noise figure as

$$F = F_{min} + R_n R_s \left|G_{opt} - G_s\right|^2 \tag{11.25}$$

which clearly shows that the rate of deviation from optimal noise figure is determined by R_n. If a two-port has a small value of R_n, then we can be sloppy and sacrifice the noise match for gain. If R_n is large, though, we have to pay careful attention to the noise match. Most software packages (Spectre, ADS) will plot Y_{opt} and F_{min} as a function of frequency, allowing the designer to choose the right match for a given bias point.

Transistor Device Selection

Previously we found the equivalent noise generators for a BJT

$$\overline{v_n^2} = \overline{v_{r_b}^2} + \frac{\overline{i_c^2}}{g_m^2} = 4kTBr_b + \frac{2qI_C B}{g_m^2} \tag{11.26}$$

$$\overline{i_n^2} = \overline{i_b^2} \tag{11.27}$$

The noise figure is therefore

$$F = 1 + \frac{4kTr_b + \frac{2qI_C}{g_m^2}}{4kTR_s} + \frac{2qI_C R_s^2}{\beta 4kTR_s} = 1 + \frac{r_b}{R_s} + \frac{1}{2g_m R_s} + \frac{g_m R_s}{2\beta} \tag{11.28}$$

According to the above expression, we can choose an optimal value of $g_m R_s$ to minimize the noise. But the second term r_b/R_s is fixed for a given transistor dimension.

The transistor base resistance r_b should be smaller than the source. This resistance can be scaled down by increasing the area of the device. This has an additional benefit of lowering the current density which delays the onset of the Kirk Effect. But a larger device means more capacitive parasitics. We can thus see that BJT transistor sizing involves a compromise:

- The transconductance depends only on I_C and not the size (unlike a FET).
- The charge storage effects and f_T only depend on the base transit time, a fixed vertical dimension.
- A smaller device has smaller junction area but can only handle a given current density before Kirk effect reduces performance.
- A larger device has smaller base resistance r_b but larger junction capacitance.

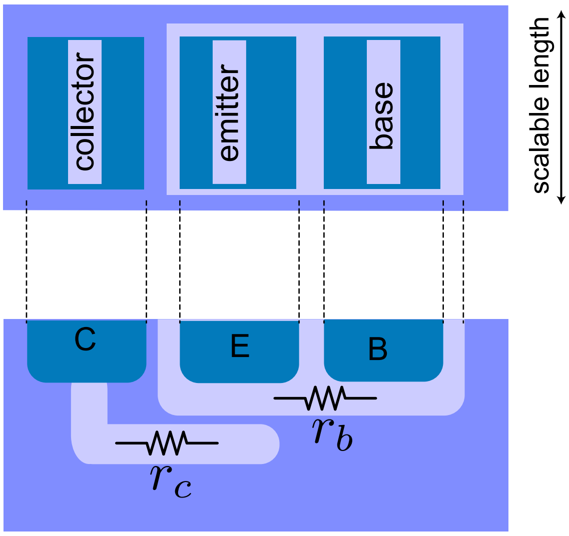


Figure 11.6: The layout of a simple bipolar junction transistor (BJT).

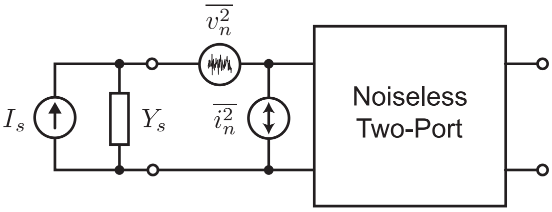


Figure 11.7: A noisy two-port is represented as a noiseless two-port with equivalent *correlated* voltage and current noise sources.

11.1.2 Two-Port Noise with Correlation

Let's calculate the noise figure of a general two-port shown in Fig. 11.7. The total input current into the device is easily calculated by superposition

$$i_i = \frac{Y_{in}}{Y_{in}+Y_s} i_s + \frac{Y_{in}}{Y_{in}+Y_s} i_n + \frac{1}{Z_{in}+Z_s} v_n \tag{11.29}$$

Here we have added noise currents and voltages as if they were deterministic signals. We need to exercise caution when computing the power entering the two-port by carefully considering the correlation among the sources. The above expression can be simplified by noting that

$$\frac{1}{Z_{in}+Z_s} v_n = \frac{Y_{in}}{Y_s+Y_{in}} Y_s v_n \tag{11.30}$$

The total noise current is thus given by

$$i_i = \frac{Y_{in}}{Y_{in}+Y_s}(i_s + i_n + Y_s v_n) \tag{11.31}$$

Now the total input RMS power is proportional to $\overline{i_i^2}\Re(Z_{in})$, and the ratio of powers can be written simply

$$F = \frac{\overline{i_i^2}}{\overline{i_i^2}\Big|_{i_n=0,v_n=0}} = 1 + \frac{\overline{(i_n+Y_s v_n)^2}}{\overline{i_s^2}} \tag{11.32}$$

The last equality follows from the fact that the source noise is independent of the two-port noise. It's fruitful to now express the input current i_n as a sum of two part, a part uncorrelated from v_n called i_u and a part that is correlated i_c

$$i_n = i_u + Y_C v_n \tag{11.33}$$

The admittance Y_C is called the correlation admittance. With the above substitution, we have

$$= 1 + \frac{\overline{(i_u + (Y_C + Y_s)v_n)^2}}{\overline{i_s^2}} \tag{11.34}$$

But now by the fact that the numerator terms are independent by design gives us

$$= 1 + \frac{\overline{i_u^2} + |Y_C + Y_s|^2 \overline{v_n^2}}{\overline{i_s^2}} \tag{11.35}$$

Let $v_n = 4kTR_n$ and $i_u = 4kTG_n$ so that we can express the above in the following form

$$F = 1 + \frac{G_n}{G_s} + \frac{R_n}{G_s}|Y_C + Y_s|^2 \tag{11.36}$$

$$= 1 + \frac{G_n}{G_s} + \frac{R_n}{G_s}((G_C + G_s)^2 + (B_C + B_s)^2) \tag{11.37}$$

Noise Optimum Source Admittance Revisited

For the general case with correlation, the optimum source impedance is easily found by taking partials of F. In particular note that the minimum F as a function of B_s is given by $-B_C$. In terms of G_s

$$\frac{\partial F}{\partial G_s} = -\frac{G_u}{G_s^2} + \frac{R_n}{G_s} 2(G_C + G_s) - \frac{(G_C + G_s)^2 R_n}{G_s^2} = 0 \tag{11.38}$$

Just as before, the solution exists can be shown to correspond to a minimum point

$$G_{s,opt} = \sqrt{G_C^2 + \frac{G_u}{R_n}} \tag{11.39}$$

$$B_{s,opt} = -B_C \tag{11.40}$$

Setting Y_c to zero, our result agrees with our previous calculation. Just as before, in general $Y_{S,opt} \neq Y_{in}^*$ and thus the noise match does not correspond to the gain match, and thus a compromise is necessary to find the best performance.

11.1.3 Scattering Parameter Representation

We have an important result from our two-port network analysis of noise. For any two-port, the noise is described by four parameters, which can be expressed as

$$F = F_{min} + \frac{R_n}{G_s}|Y_s - Y_{s,opt}|^2 \tag{11.41}$$

F_{min} is the minimum achievable noise figure when the source admittance is set to $Y_{s,opt}$, and deviates from this value according to the scale factor R_n.

Unfortunately, we are not always free to choose $Y_{s,opt}$ as our source admittance due to other conflicting factors (gain, stability). So how can we select Y_s to get the best overall performance? A powerful way to visualize the trade-off in noise peformance versus source impedance is to analyze the problem in the scattering parameter domain and plot the noise figure as a function of the source reflection coefficient as contours on the Smith Chart.

Source Reflection Representation

Our ultimate goal is to plot the noise figure on the Smith Chart so that we can compare the noise performance to gain circles. It's desirable to represent the noise in terms of the source reflection coefficient

$$|Y_S - Y_{opt}|^2 = \left| \frac{1-\Gamma_S}{1+\Gamma_S} - \frac{1-\Gamma_{opt}}{1+\Gamma_{opt}} \right|^2 Y_0^2 \tag{11.42}$$

$$= \left| \frac{(1-\Gamma_S)(1+\Gamma_{opt}) - (1-\Gamma_{opt})(1+\Gamma_S)}{(1+\Gamma_S)(1+\Gamma_{opt})} \right|^2 Y_0^2 \tag{11.43}$$

Simplifying the numerator, we have

$$= 4Y_0^2 \frac{|\Gamma_{opt} - \Gamma_S|^2}{|1+\Gamma_S|^2 \, |1+\Gamma_{opt}|^2} \tag{11.44}$$

$$G_S = \Re(Y_S) = \frac{1}{2}(Y_S + Y_S^*) = \frac{1}{2} Y_0 \left(\frac{1-\Gamma_S}{1+\Gamma_S} + \frac{1-\Gamma_S^*}{1+\Gamma_S^*} \right) \tag{11.45}$$

$$= \frac{Y_0}{2} \frac{(1-\Gamma_S)(1+\Gamma_S^*) + (1-\Gamma_S^*)(1+\Gamma_S)}{|1+\Gamma_S|^2} = Y_0 \frac{1-|\Gamma_S|^2}{|1+\Gamma_S|^2} \tag{11.46}$$

With these substitutions, we have

$$F = F_{min} + \frac{4R_n Y_0 |\Gamma_{opt} - \Gamma_S|^2}{(1-|\Gamma_S|^2)\,|1+\Gamma_{opt}|^2} \tag{11.47}$$

Does this look like a circle? Not yet, unless you have a very well trained eye. We will next show that for a fixed value of F, this is an equation for a circle in the Γ plane. Collect all terms that are constant for fixed F

$$N = \frac{F - F_{min}}{4R_n Y_0} |1+\Gamma_{opt}|^2 \tag{11.48}$$

With this definition, the equation simplifies to

$$N = \frac{|\Gamma_S - \Gamma_{opt}|^2}{1-|\Gamma_S|^2} \tag{11.49}$$

$$N(1-|\Gamma_S|^2) = (\Gamma_S - \Gamma_{opt})(\Gamma_S^* - \Gamma_{opt}^*) \tag{11.50}$$

$$N(1-|\Gamma_S|^2) = \Gamma_S^2 - (\Gamma_{opt}\Gamma_S^* + \Gamma_{opt}^*\Gamma_S) - |\Gamma_{opt}|^2 \tag{11.51}$$

$$|\Gamma_S|^2(N+1) - (\Gamma_{opt}\Gamma_S^* + \Gamma_{opt}^*\Gamma_S) - |\Gamma_{opt}|^2 - N = 0 \tag{11.52}$$

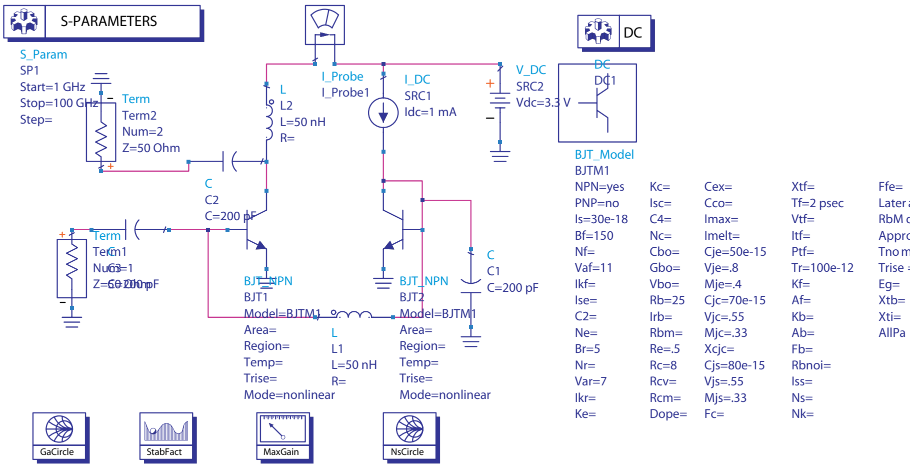


Figure 11.8: An ADS test setup to analyze the gain and noise performance of a BJT amplifier.

Completing the square,

$$|\Gamma_S|^2 - \frac{1}{N+1}(\Gamma_{opt}\Gamma_S^* + \Gamma_{opt}^*\Gamma_S) + \frac{|\Gamma_{opt}|^2}{(N+1)^2} = \frac{|\Gamma_{opt}|^2 + N}{N+1} + \frac{|\Gamma_{opt}|^2}{(N+1)^2} \tag{11.53}$$

which is rewritten as

$$\left|\Gamma_S - \frac{\Gamma_{opt}}{N+1}\right| = \frac{\sqrt{N(N+1-|\Gamma_{opt}|^2)}}{N+1} \tag{11.54}$$

We are now in a position to say that indeed, the equation is a circle. The noise circles are centered at $\Gamma_{opt}/(N+1)$ with radius given by the right hand side. Note that, as expected, the optimum noise figure is a point centered at Γ_{opt} with radius zero ($N = 0$).

Example 32:BJT Amplifier Example

A BJT device in common emitter configuration is shown in Fig. 11.8. The simulation setup will calculate the S parameters, noise figure, and available gain circles.

A plot of the maximum stable gain (MSG), $NF_{50\,\Omega}$, and NF_{min} is shown in Fig. 11.9. Note the f_{max} is around 30GHz with about 15.7dB stable gain at 5GHz. The minimum achievable noise figure is about 1.25dB but the 50 Ω noise figure is considerably higher.

What is the best achievable noise/gain trade-off? By plotting the available gain circles G_A along with the noise figure circles (Fig. 11.10, we can choose an appropriate point to achieve a reasonable trade-off. For instance, the $G_A = 15$dB circle intersects the $NF = 1.5$dB circle when the source impedance is $Z_S = Z_0(2.4 + j0.6)$.

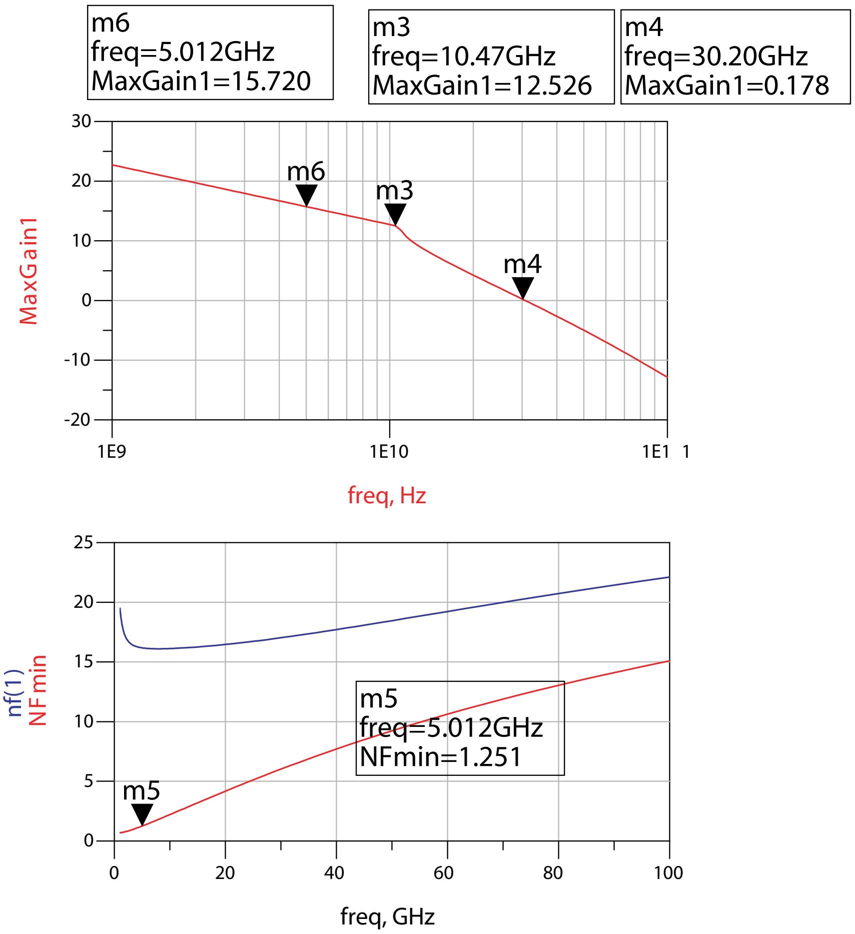
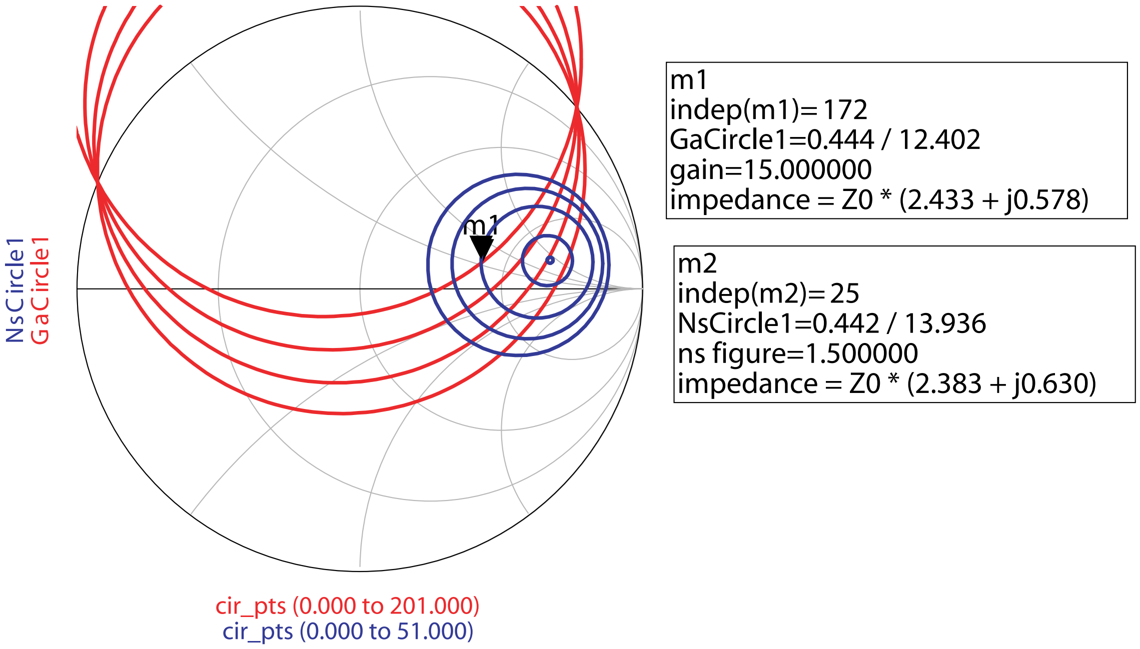


Figure 11.9: Simulated G_{max} maximum gain for the device and the minimum achievable noise figure (and noise figure for a 50Ω source.



m1
indep(m1)= 172
GaCircle1=0.444 / 12.402
gain=15.000000
impedance = Z0 * (2.433 + j0.578)

m2
indep(m2)= 25
NsCircle1=0.442 / 13.936
ns figure=1.500000
impedance = Z0 * (2.383 + j0.630)

Figure 11.10: Simulated Noise and Gain circles.

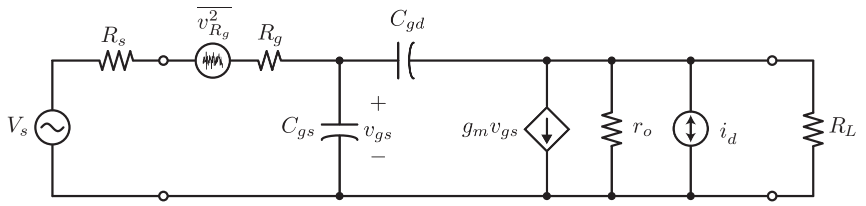


Figure 11.11: FET common-source (CS) amplifier noise equivalent circuit model.

11.2 FET Common Source Amplifier

Let's analyze the noise performance of a FET Common Source (CS) amplifier, as shown in Fig. 11.11. The mdoel contains the following noise sources:

R_s	$\overline{v_{v_s}^2} = 4kTBR_s$
R_g	$\overline{v_g^2} = 4kTBR_g$
R_{ch}	$\overline{i_d^2} = 4kTBg_{d0}\gamma B$
R_L	$\overline{i_L^2} = 4kTBG_L$

11.2.1 CS Noise at Low Frequencies

Summing all the noise at the output (assume low frequency)

$$\overline{i_o^2} = \overline{i_d^2} + \overline{i_L^2} + (\overline{v_g^2} + \overline{v_s^2})g_m^2 \tag{11.55}$$

Which results in the noise figure

$$F = 1 + \frac{\overline{v_g^2}}{\overline{v_s^2}} + \frac{\overline{i_d^2} + \overline{i_L^2}}{g_m^2 \overline{v_s^2}} \tag{11.56}$$

$$= 1 + \frac{R_g}{R_s} + \frac{g_{d0}\gamma + G_L}{R_s g_m^2} \tag{11.57}$$

Assume $g_m = g_{d0}$ (long channel approximation)

$$= 1 + \frac{R_g}{R_s} + \frac{\gamma}{g_m R_s} + \frac{G_L G_S}{g_m^2} \tag{11.58}$$

If we make g_m sufficiently large, the gate resistance will dominate the noise. The gate resistance has two components, the physical gate resistance and the induced channel resistance

$$R_G = R_{poly} + \delta R_{ch} = \frac{1}{3}\frac{W}{L}R_{\square} + \frac{1}{5}\frac{1}{g_m} \tag{11.59}$$

The factors $1/3$ and $1/5$ come from a distributed analysis of the channel noise. They are valid for single-sided gate contacts. To reduce the gate resistance, a multi-finger layout approach (Fig. 11.12) is commonly adopted. As a bonus, the junction capacitance is reduced due to the junction sharing.

11.2.2 CS Noise at Medium Frequencies

If we repeat the calculation at medium frequencies (Fig. 11.13), ignoring C_{gd}, we simply need to input refer the drain noise taking into account the frequency dependence of G_m

$$G_m = g_m \frac{1/(j\omega C_{gs})}{1/(j\omega C_{gs}) + R_s + R_g} \tag{11.60}$$

$$= \frac{g_m}{1 + j\omega C_{gs}(R_s + R_g)} \tag{11.61}$$

The drain noise is input referred by the magnitude squared

$$|G_m|^{-2} = g_m^{-2}(1 + \omega^2 C_{gs}^2 (R_s + R_g)^2) \tag{11.62}$$

So the noise figure is simply given by (neglect the noise of R_L)

$$F = 1 + \frac{R_g}{R_s} + \frac{\gamma}{\alpha}\left(1 + \omega^2 C_{gs}^2 (R_S + R_g)^2\right) \tag{11.63}$$

Assume that $R_s \gg R_g$ (good layout). The medium frequency noise is given by

$$F_\infty = 1 + \frac{\gamma}{\alpha}\frac{\omega^2 C_{gs}^2 R_s^2}{g_m R_S} = 1 + \frac{\gamma}{\alpha}\left(\frac{\omega}{\omega_T}\right)^2 g_m R_s \tag{11.64}$$

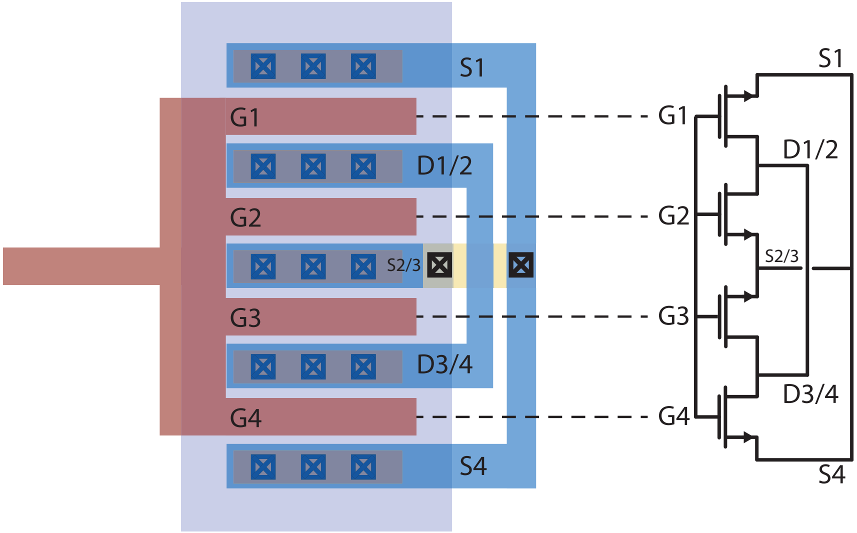


Figure 11.12: FET multi-finger device layout to minimize gate resistance and junction capacitance.

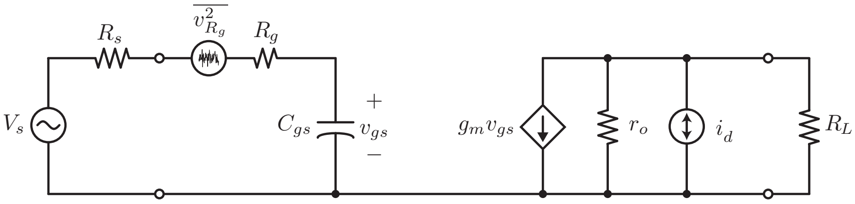


Figure 11.13: FET common-source amplifier model at medium frequencies (neglecting C_{gd}).

11.2.3 MOS Amplifier Current Gain

Let's recalculate the MOS amp noise figure (quickly). Note that the current gain of the MOS amp is given by

$$i_o = g_m v_1 = g_m \frac{v_s}{R_s + R_g + \frac{1}{j\omega C_{gs}}} \left(\frac{1}{j\omega C_{gs}} \right) \tag{11.65}$$

$$= v_s \frac{g_m}{1 + j\omega C_{gs}(R_s + R_g)} \approx v_s \frac{g_m}{j\omega C_{gs}(R_s + R_g)} \tag{11.66}$$

This can be rewritten as $i_o = G_m v_s$, where

$$G_m = -j \frac{\omega_T}{\omega} \frac{1}{R_s + R_g} \tag{11.67}$$

This facilitates the noise calculations since the total noise is given by

$$\overline{i_{o,T}^2} = G_m^2 (\overline{v_g^2} + \overline{v_s^2}) + \overline{i_d^2} \tag{11.68}$$

And the noise figure is easily computed

$$F = 1 + \frac{\overline{v_g^2}}{\overline{v_s^2}} + \frac{\overline{i_d^2}}{G_m^2 \overline{v_s^2}} \tag{11.69}$$

Substitution of the the various noise sources leads to

$$F = 1 + \frac{R_g}{R_s} + \frac{\gamma g_m}{R_s} \left(\frac{\omega}{\omega_T} \right)^2 (R_s + R_g)^2 \tag{11.70}$$

Assume that $R_s \gg R_g$ to get

$$F = 1 + \frac{R_g}{R_s} + \gamma \left(\frac{\omega}{\omega_T} \right)^2 g_m R_s \tag{11.71}$$

As noted earlier, this expression contains both the channel noise and the gate induced noise. If we assume that $R_g = R_{poly} + \frac{1}{5g_m}$, and the noise is *independent* from the drain thermal noise, we get a very good approximation to the actual noise without using correlated noise sources.

11.3 Minimum Noise for MOS Amp

Let's find the optimal value of R_s

$$\frac{\partial F}{\partial R_s} = -\frac{R_g}{R_s^2} + \gamma \left(\frac{\omega}{\omega_T} \right)^2 g_m = 0 \tag{11.72}$$

or

$$\frac{R_g}{R_s^2} = \gamma \left(\frac{\omega}{\omega_T} \right)^2 g_m \tag{11.73}$$

Solving

$$R_{s,opt} = \left(\frac{\omega_T}{\omega} \right) \sqrt{\frac{R_g}{\gamma g_m}} \tag{11.74}$$

We now have (after simplification)

$$F_{min} = 1 + 2\left(\frac{\omega}{\omega_T}\right)\sqrt{g_m R_g \gamma} \tag{11.75}$$

Example 33:MOS Amp Example

Let's find $R_{s,opt}$ for a typical amplifier. Say $f_T = 75\,\mathrm{GHz}$, $f = 5\,\mathrm{GHz}$, and $\gamma = 2$. Also suppose that by proper layout R_{poly} is very small. The intrinsic gate resistance is given by

$$R_g = R_{poly} + \frac{1}{5g_m} \approx \frac{1}{5g_m} \tag{11.76}$$

To make the noise contribution from this term 0.1 requires that

$$\frac{R_g}{R_s} = 0.1 \tag{11.77}$$

$$\frac{1}{5g_m R_s} = 0.1 \tag{11.78}$$

$$5g_m R_s = 10 \tag{11.79}$$

$$g_m = \frac{10}{5 \times 50\,\Omega} = \frac{1}{25}\,\mathrm{S} = 40\,\mathrm{mS} \tag{11.80}$$

Note that for $V_{gs} - V_T = 800\,\mathrm{mV}$ (set by the desired high f_T), the required current is fairly hefty

$$g_m = \frac{2I_{ds}}{V_{gs} - V_T} = 40\,\mathrm{mS} \tag{11.81}$$

and

$$I_{ds} = 40\,\mathrm{mS} \times 800\,\mathrm{mV} \times \frac{1}{2} = 16\,\mathrm{mA} \tag{11.82}$$

The optimum source resistance is given by

$$R_{s,opt} = \frac{f_T}{f}\sqrt{\frac{R_g}{\gamma g_m}} = 15\sqrt{\frac{5 \cdot 25}{2}} \approx 119\Omega \tag{11.83}$$

which gives a very low noise figure of

$$F_{min} = 1 + 2\frac{f}{f_T}\sqrt{g_m R_g \gamma} = 1 + \frac{2}{15}\sqrt{5 \times 2/25} = 1.08 \tag{11.84}$$

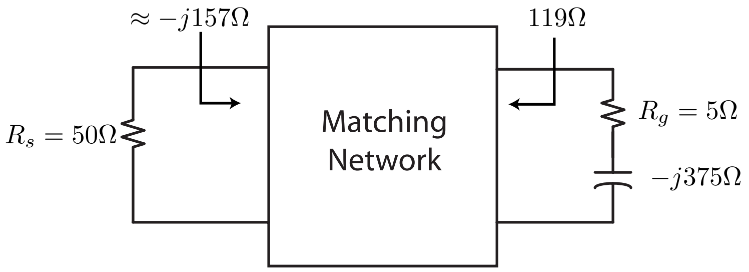


Figure 11.14: A noise matching scheme that converts the source impedance to the optimum source impedance as shown.

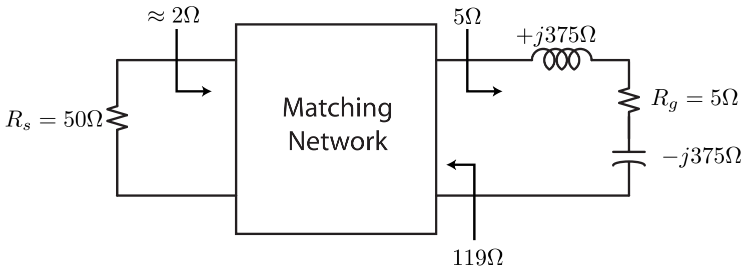


Figure 11.15: A noise matching scheme converts the source impedance to the optimum source impedance as shown and tunes out the large reactive part of the FET input impedance using a series inductor.

The LNA MOS amplifier example was able to achieve a very low noise figure of .35 dB. In practice, though, it'll be difficult to get this low of a noise figure and get useful gain with the simple common source. To see why, note that $C_{gs} \approx g_m/\omega_T = 85\,\text{fF}$. The input impedance of the FET is given by

$$Z_i = R_g + \frac{1}{j\omega C_{gs}} = R_g - j\frac{\omega_T}{\omega g_m} \approx 5\Omega - j375\Omega \tag{11.85}$$

This highlights the need for matching.

11.4 Matching for Noise and Gain

11.4.1 Matching Option 1

As shown in Fig. 11.14, let's say we don't match the input impedance. Simply use a matching network to multiply the 50Ω source up to 119Ω. This means that the source (antenna) will see a termination that is $m = 119/50 = 2.38$ times smaller, or about $-j157\Omega$. This is a good for noise but a bad choice for a power match.

11.4.2 Matching Option 2

As shown in Fig. 11.15, we could use an inductor to tune out the capacitive part of the input. This will add some noise due to finite inductor Q, but for now let's ignore that contribution. Note that

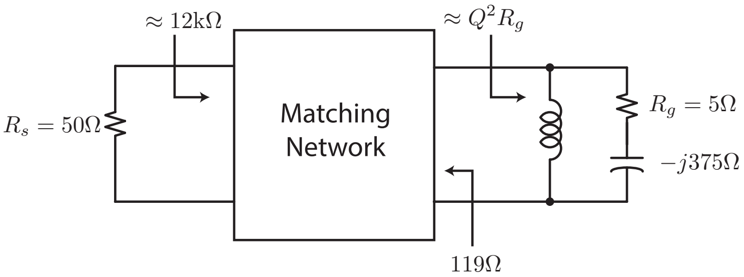


Figure 11.16: A noise matching scheme converts the source impedance to the optimum source impedance as shown and tunes out the large reactive part of the FET input impedance using a shunt inductor.

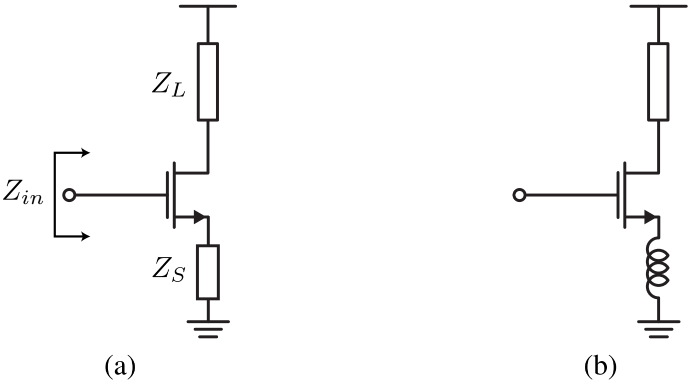


Figure 11.17: (a) A common-source amplifier with series-series feedback, or source degeneration. (b) Inductive degeneration.

the matching network will match this low 5Ω resistance down to 5Ω/2.38 ≈ 2Ω. Now the power match is even worse.

11.4.3 Matching Option 3

Instead of a series inductor, let's use a shunt inductor to resonate the input impedance, as shown in Fig. 11.16. The inductor should be connected to the DC value of V_{gs} and can double as a biasing element. But since the gate capacitance is high Q

$$Q = \frac{1}{\omega C_{gs} R_g} \approx \frac{1}{\omega C_{gs} \frac{1}{5 g_m}} = 5 \frac{f_T}{f} = 5 \times 15 = 75 \tag{11.86}$$

The input resistance is going to be $Q^2 R_g \approx 28\,\mathrm{k}\Omega$, or too big. The matching circuit will bring this "down" to about 12 kΩ, a very poor match.

11.4.4 Source/Emitter Degeneration

None of the above presented matching techniques work, particularly at low frequencies when the input Q is high. An alternative is source (or emitter for a BJT) degeneration. To understand this technique, let's first derive the input impedance looking into the gate (base) when a reactance is placed on the source (emitter) side. Connect a test current source i_x and measure the voltage (Fig. 11.17a).

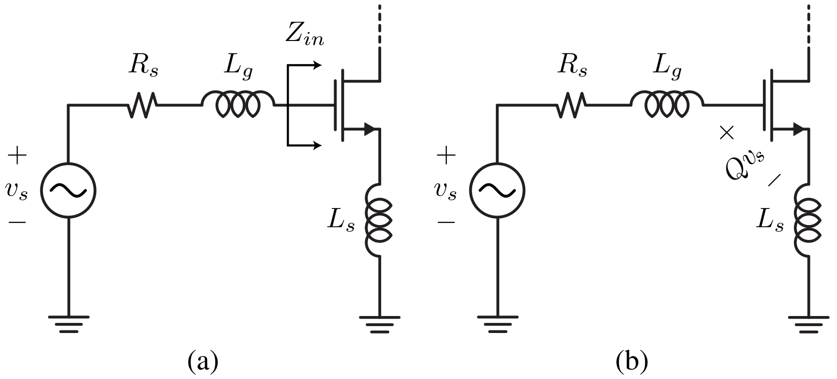


Figure 11.18: (a) The input impedance of an inductively degeneration common source amplifier. (b) Q boosting of the series RLC network increases the voltage on C_{gs}.

The voltage at the input of the amplifier is given by

$$v_x = i_x Z_{gs} + (i_x + g_m Z_{gs} i_x) Z_s \tag{11.87}$$

$$Z_{in} = Z_s + Z_{gs} + \underbrace{g_m Z_{gs} Z_s}_{\text{due to feedback}} \tag{11.88}$$

Let's assume that Z_s is reactive (zero noise)

$$g_m Z_{gs} Z_s = g_m \frac{1}{j\omega C_{gs}} jX = \frac{g_m X}{\omega C_{gs}} \tag{11.89}$$

which produces a purely passive input resistance if $X > 0$

11.5 Inductive Degeneration

The reactive feedback from an inductor (Fig. 11.17b) produces a broadband programmable real input impedance that can simplify matching (or even eliminate it altogether).

$$\Re(Z_{in}) = \frac{g_m L}{C_{gs}} \approx \omega_T L \tag{11.90}$$

We thus select L by $L = \frac{R_s}{\omega_T}$. If this value of L is impractical, we can artificially reduce ω_T by inserting a capacitor in shunt with C_{gs}.

The input impedance of the FET with inductive degeneration is given by

$$Z_{in} = j\omega L_s + \frac{1}{j\omega C_{gs}} + \omega_T L_s = j\omega L_s + \frac{1}{j\omega C_{gs}} + R_s \tag{11.91}$$

The input impedance behaves like a series RLC circuit. We need to tune the resonant frequency of the series circuit to align with the operating frequency. This can be done by adding gate inductance L_g, as shown in Fig. 11.18a.

Recall that in a resonant circuit, the voltage across the reactive elements is Q times larger than the voltage across the resistor (see Fig. 11.18b). At resonance, the voltage across the resistors is simply v_s, so we have

$$v_{gs} = Q \times v_s \tag{11.92}$$

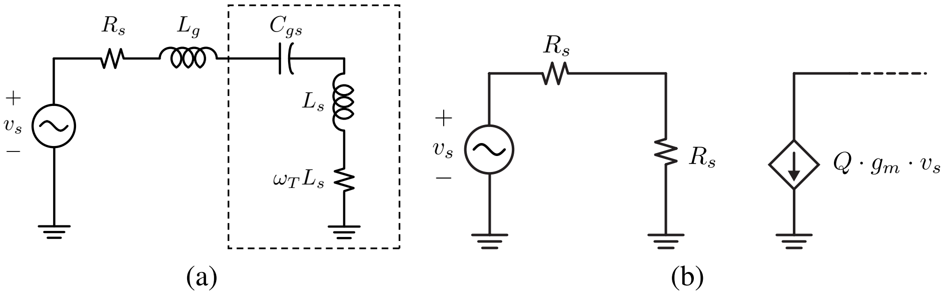


Figure 11.19: (a) Equivalent circuit for the input impedance of the inductively degeneration amplifier. (b) Equivalent circuit at resonance.

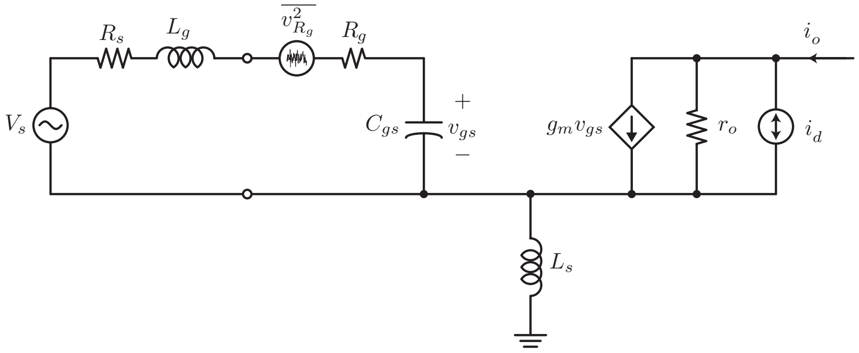


Figure 11.20: Noise equivalent circuit for inductively degenerated common-source amplifier.

$$i_d = g_m v_{gs} = Q \times g_m v_s = G_m v_s \tag{11.93}$$

$$Q = \frac{1}{\omega_0 C_{gs} 2R_s} \tag{11.94}$$

$$G_m = Qg_m = \frac{g_m}{\omega_0 C_{gs} 2R_s} = \left(\frac{\omega_T}{\omega_0}\right)\frac{1}{2R_s} \tag{11.95}$$

11.5.1 Equivalent Circuit at Resonance

From the source, the amplifier input (ignoring C_{gd}) is equivalent to the following circuit shown in Fig. 11.19a. At resonance, the circuit simplifies to Fig. 11.19b.

11.5.2 Noise Figure for Inductive Degeneration

It's fairly easy to calculate the noise for the case with inductive degeneration shown in Fig 11.20. Simply observe that the input generators ($\overline{v_s^2}$ and $\overline{v_g^2}$) see a gain of G_m^2 to the output. The drain noise $\overline{i_d^2}$, though, requires a careful analysis. Since $\overline{i_d^2}$ flows partly into the source of the device, it activates the g_m of the transistor which produces a correlated noise in shunt with $\overline{i_d^2}$.

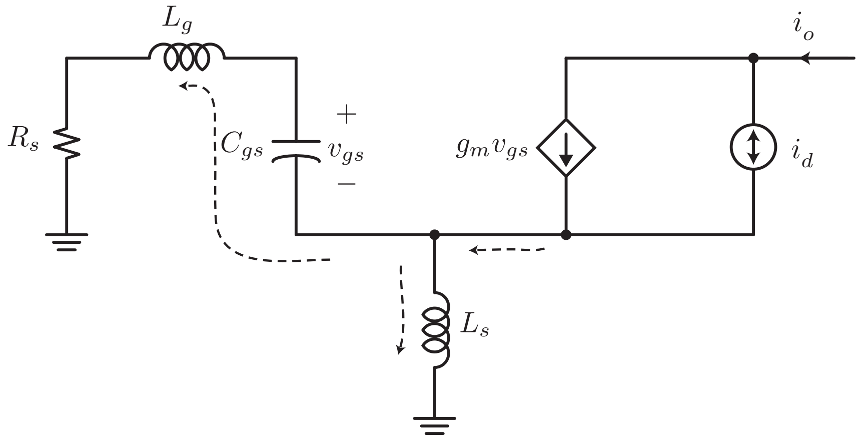


Figure 11.21: The drain noise of the FET is divided between the inductive degeneration and the input network, which actives the g_m of the transistor.

The equivalent circuit of Fig. 11.21, shows that the noise component flowing into the source is given by the current divider

$$v_\pi = -(g_m v_\pi + i_d) \times \frac{j\omega L_s}{j\omega L_s + \frac{1}{j\omega C_{gs}} + j\omega L_g + R_s} \times \frac{1}{j\omega C_{gs}} \tag{11.96}$$

The denominator simplifies to R_s at resonance, so we have

$$v_\pi = -(g_m v_\pi + i_d) \times \frac{j\omega L_s}{R_s} \frac{1}{j\omega C_{gs}} \tag{11.97}$$

$$= -(g_m v_\pi + i_d) \times \frac{L_s}{C_{gs} R_s} \tag{11.98}$$

$$v_\pi \left(1 + \frac{g_m L_s}{C_{gs} R_s}\right) = -i_d \frac{L_s}{C_{gs} R_s} \tag{11.99}$$

But $\omega_T L_s = R_s$, so we have

$$2 v_\pi = -i_d \frac{L_s}{C_{gs} R_s} \tag{11.100}$$

or

$$g_m v_\pi = -\frac{i_d}{2} \frac{g_m L_s}{C_{gs} R_s} = -\frac{i_d}{2} \tag{11.101}$$

So we see that only $1/4$ of the drain noise flows into the output. The total output noise is therefore

$$\overline{i_{o,T}^2} = G_m^2(\overline{v_s^2} + \overline{v_g^2}) + \frac{1}{4}\overline{i_d^2} \tag{11.102}$$

so the noise factor is

$$F = 1 + \frac{\overline{v_g^2}}{\overline{v_s^2}} + \frac{\overline{i_d^2}}{4\overline{v_s^2} G_m^2} \tag{11.103}$$

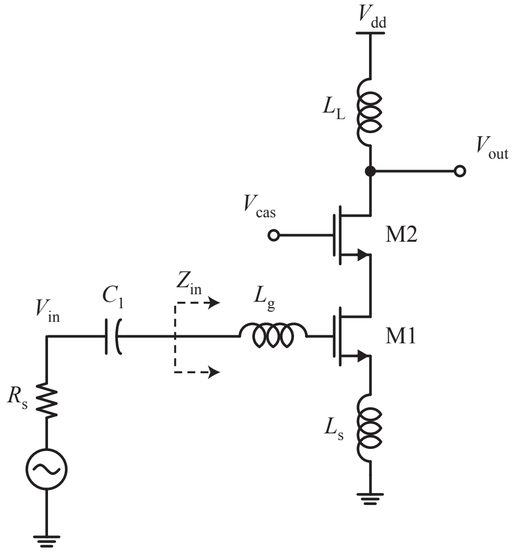


Figure 11.22: A cascode LNA with inductive degeneration has many benefits (higher gain and better isolation) without incurring much more noise.

Substitute as before and we have

$$F = 1 + \frac{R_g}{R_s} + \frac{\gamma g_m (2R_s)^2}{4R_s} \left(\frac{\omega}{\omega_t}\right)^2 \tag{11.104}$$

Note that the noise figure is the same as the common source amplifier

$$F = 1 + \frac{R_g}{R_s} + \gamma g_m R_s \left(\frac{\omega}{\omega_t}\right)^2 \tag{11.105}$$

The inductive degeneration did not raise the noise, so the minimum noise figure F_{min} is the same. The advantage is that the input impedance is now real and programmable ($\omega_T L_s$). By proper sizing, it's possible to obtain a noise and power match. This is in contrast to all the previously mentioned matching techniqeus which failed to provide a decent power match under noise matching.

11.5.3 Cascode LNA

It's very common to use a cascode device instead of a common source device. This simplifies matching since the cascode device is nearly unilateral. Let's show that the cascode device adds virtually no noise at low/medium frequencies.

As shown in Fig. 11.23, the noise contribution from the cascode is small due to the degeneration. For simplicity assume the transistor degeneration is r_o. Then most of the drain noise current will flow into C_{gs} at high frequency

$$v_\pi = (g_m v_\pi + i_d)\frac{1}{j\omega C_{gs}} \tag{11.106}$$

or

$$v_\pi(j\omega C_{gs} - g_m) = i_d \tag{11.107}$$

since

$$g_m v_\pi = \frac{-g_m}{g_m - j\omega C_{gs}} i_d = \frac{-1}{1 - j\frac{\omega}{\omega_T}} i_d \approx -i_d \tag{11.108}$$

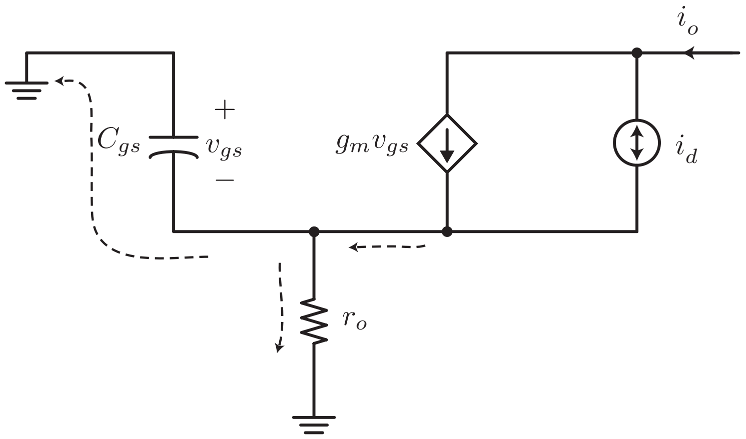


Figure 11.23: Noise equivalent circuit for the cascode device. The input g_m device is modeled by its output resistance r_o.

A similar calculation shows that at low frequency, the noise into r_o produces an output current noise of

$$(i_d + g_m v_\pi) r_o = -v_\pi \tag{11.109}$$

$$i_d r_o = -v_\pi - g_m r_o v_\pi = -(1 + g_m r_o) v_\pi \tag{11.110}$$

$$v_\pi = \frac{-r_o}{1 + g_m r_o} i_d \tag{11.111}$$

$$g_m v_\pi = \frac{-g_m r_o}{1 + g_m r_o} i_d \approx -i_d \tag{11.112}$$

The total current noise is therefore

$$\left(1 - \frac{-g_m r_o}{1 + g_m r_o}\right) i_d = \left(\frac{1}{1 + g_m r_o}\right) i_d \approx 0 \tag{11.113}$$

11.6 LNA Chip/Package/Board Interface

Since the LNA needs to interface to the external world, its input network must transition from the Si chip to the package and board environment, which involves “macroscopic” structures such as bondwires and package leads, shown in Fig. 11.24.

11.6.1 Bond Wire Inductance

One reason inductive degeneration is popular is because we can use package parasitics to our benefit. Some or all of L_s can be absorbed into the loop inductance (or the *partial* inductance of the bondwire). These parasitics must be absorbed into the LNA design. This requires a good model for the package and bondwires. It should be noted that the inductance of the input loop depends on the arrangement of the bondwires, and hence die size and pad locations. Many designs also require ESD protection, which manifests as increased capacitance on the pads and additional loss and noise figure degradation.

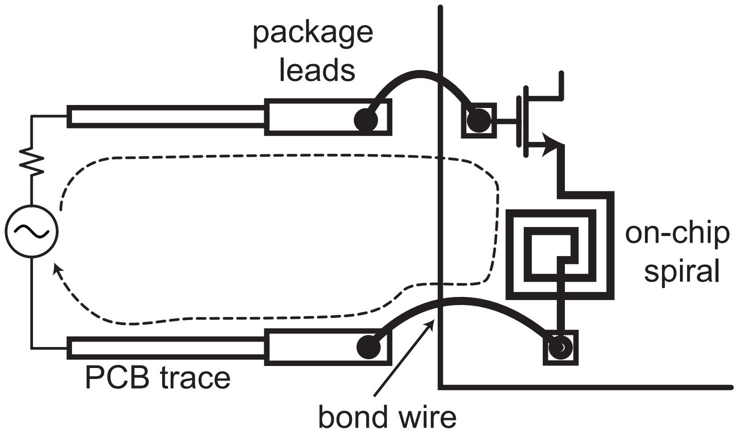


Figure 11.24: The interface between an on-chip LNA device and the external bondwire, packaging, and board parasitics.

11.6.2 Package Parasitics

Recall that a changing flux generates an emf around a circuit loop. Let

$$L = \frac{\psi}{I} \tag{11.114}$$

$$v_{emf} = \frac{d\psi}{dt} = L\frac{dI}{dt} \tag{11.115}$$

Note that in reality ψ is composed of flux from all the loops in the package, causing undesired mutual coupling to other parts of the circuit

$$v_{emf} = \frac{d(\psi_1 + \psi_2 + \psi_2 + \cdots)}{dt} = L\frac{dI_1}{dt} + M_{12}\frac{dI_2}{dt} + \cdots \tag{11.116}$$

Stability is an important consideration in LNA design since parasitic feedback paths at RF through the package, and also the substrate, can cause oscillations or instability.

11.6.3 Differential LNA Design

One undesired consequence of the package is that the parasitic inductors vary from part to part and require careful modeling and extra care to correctly implement the LNA. The advantage of a differential LNA, shown in Fig. 11.25, is that the parasitics are only on the gate side, and not on the source of the transistors. The source inductors are realized with on-chip inductors with tight process tolerances.

11.7 Closing Thoughts

By viewing the amplifier as a noisy two-port, we derived conditions to realize the lowest possible noise figure as a function of source impedance. We also showed that the a fixed noise figure maps to a circle on the Smith Chart, which is a powerful tool for analyzing the trade-off in noise versus the source impedance and matching.

If you have a discrete device, your only control knob is to vary the bias point, or perhaps explore circuit topologies that trade-off matching, gain, and noise. Typically the optimum source

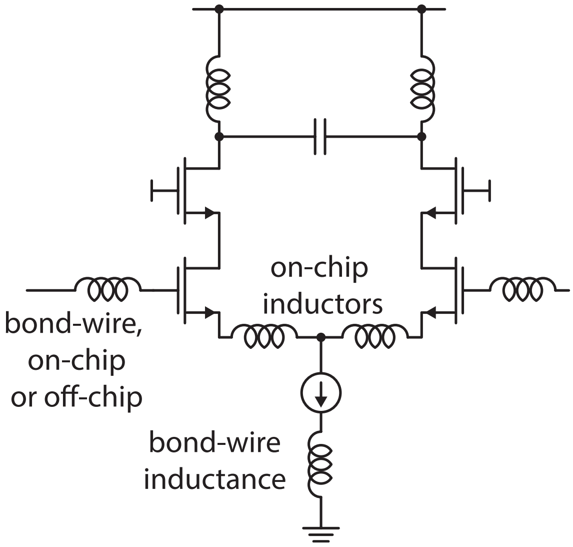
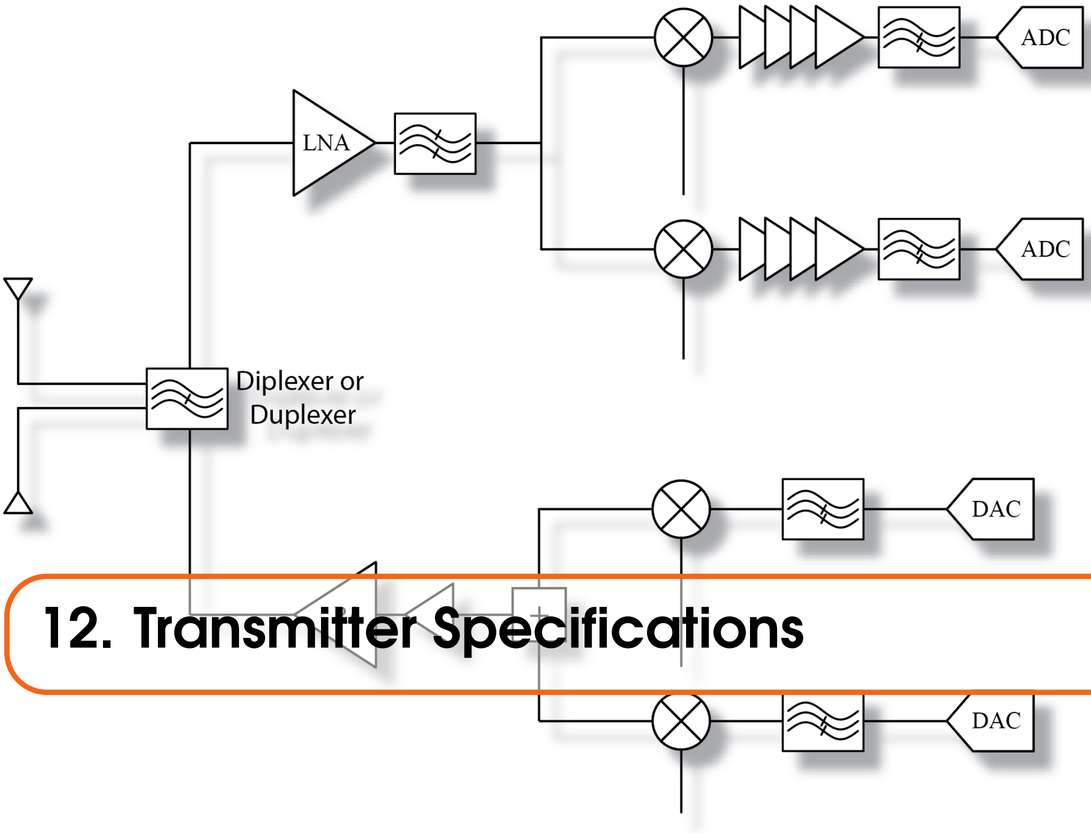


Figure 11.25: A fully differential LNA utilizes well controlled on-chip degeneration inductors.

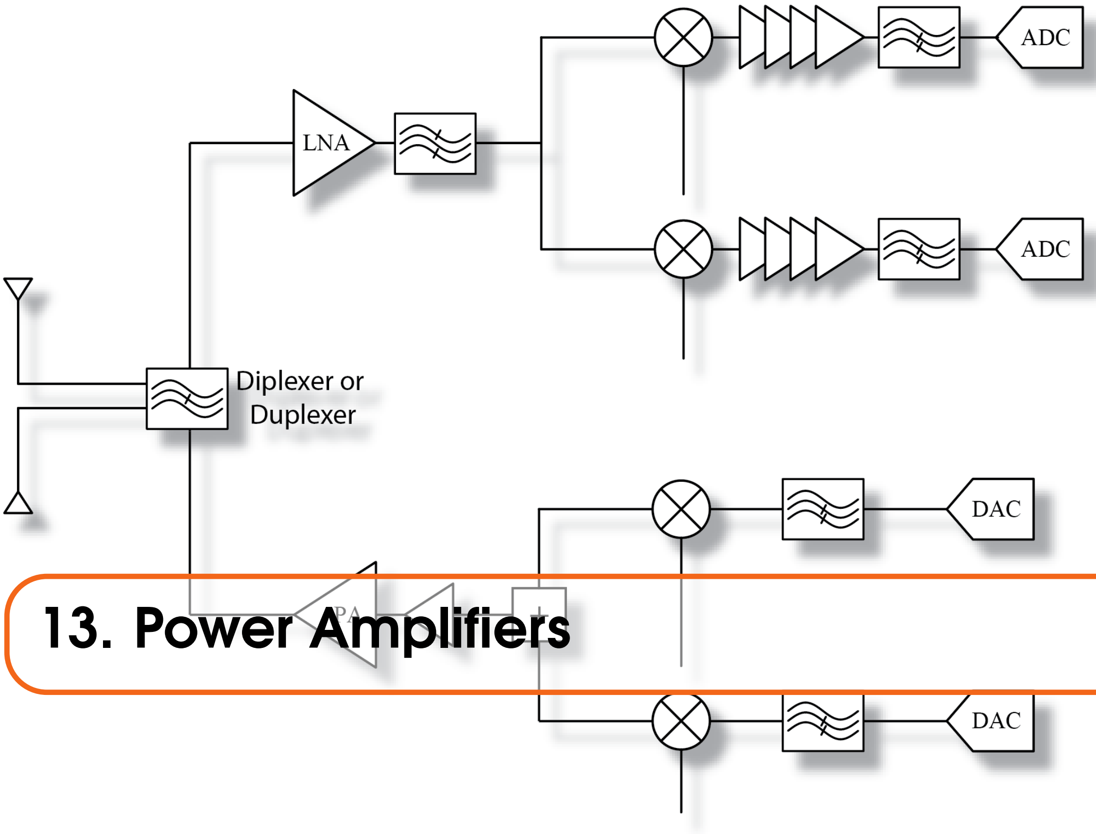
impedance for noise will not coincide with the optimum impedance for gain. The Smith Chart provides a nice way to compare gain, noise, and stability in this scenario.

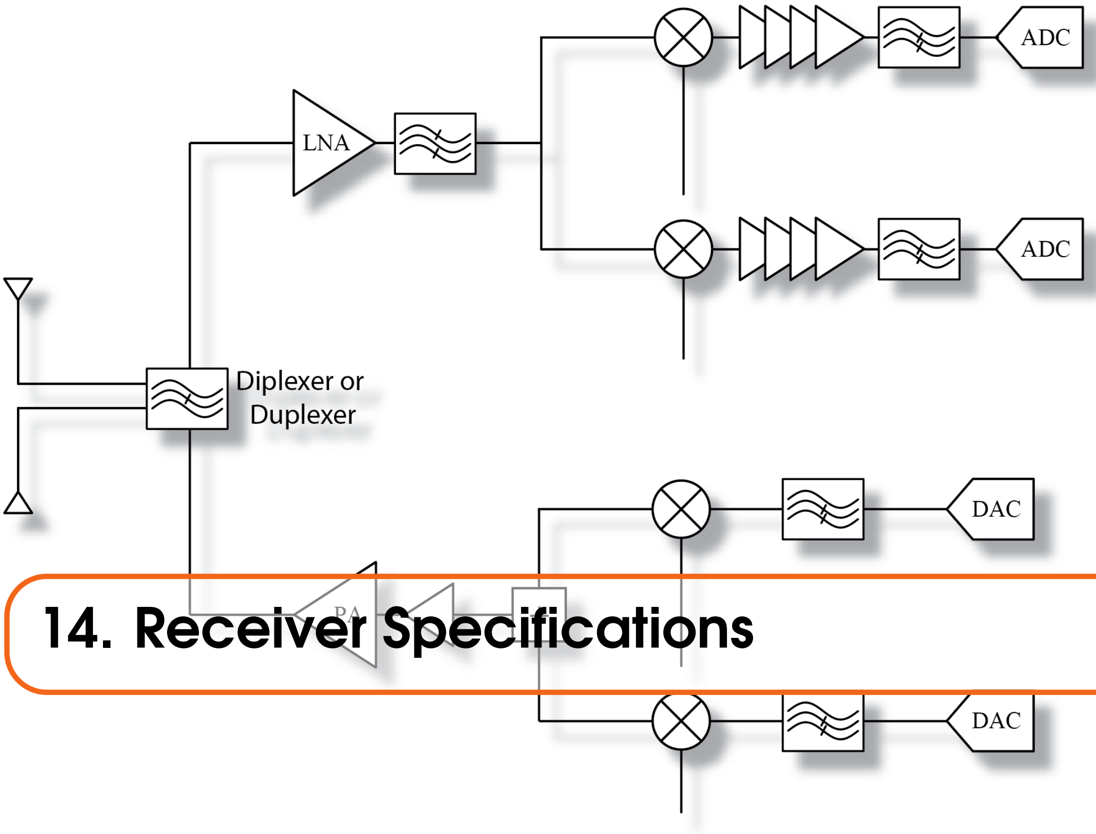
It's important to note that if you are designing an integrated circuit receiver, you have an additional degrees of freedom in choosing the device sizing and layout, all of which have a big impact on the noise performance. In fact, using this approach, you can strive to achieve simultaneous noise and gain matching.



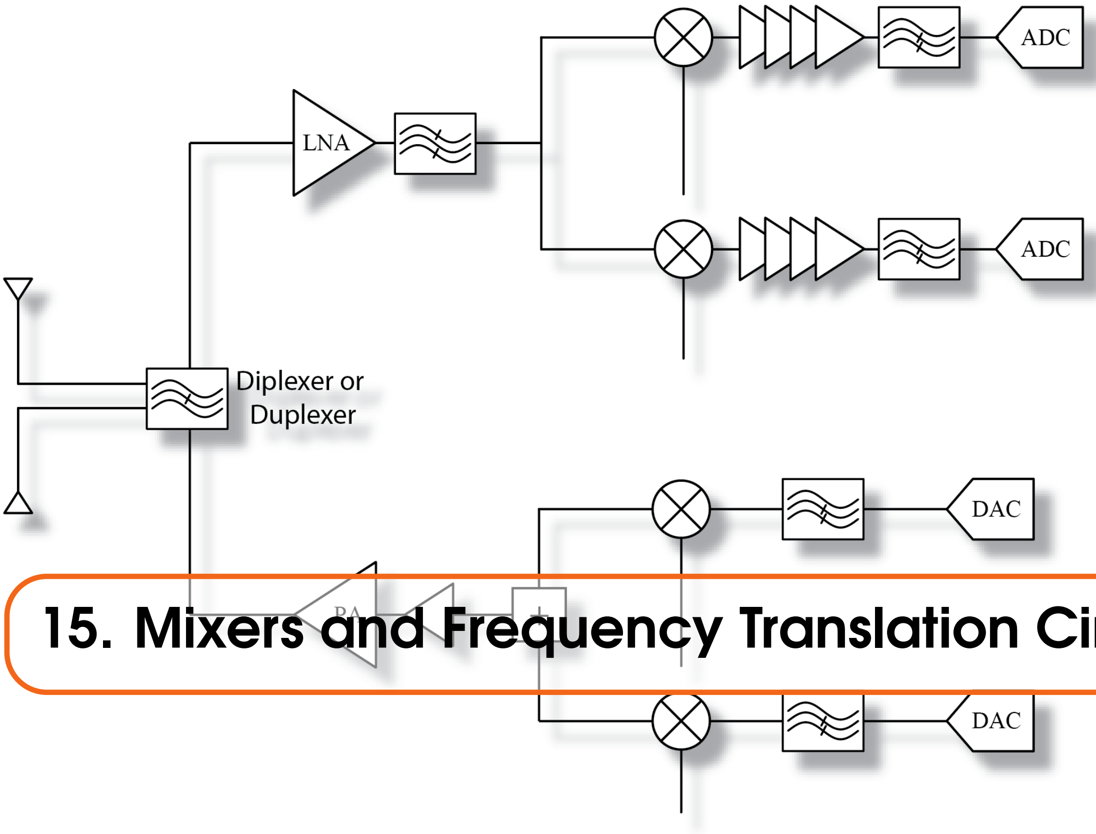
12. Transmitter Specifications

13. Power Amplifiers

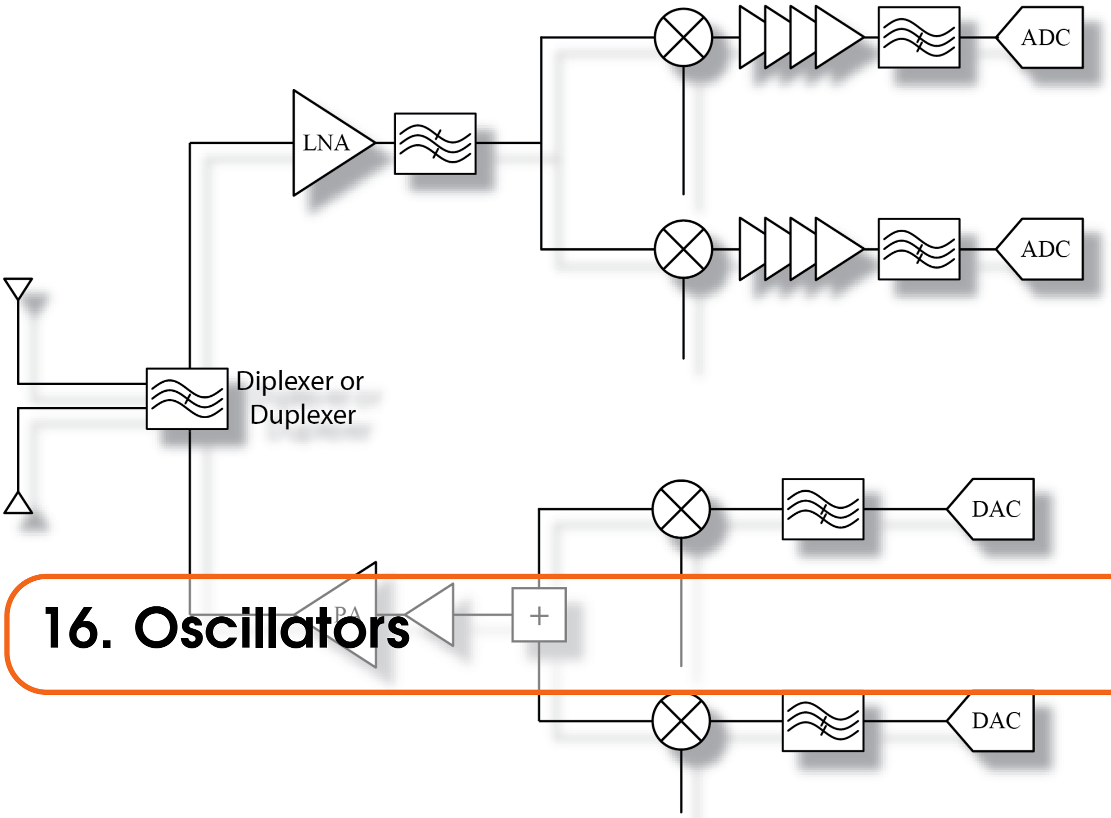




14. Receiver Specifications



15. Mixers and Frequency Translation Circuits



16. Oscillators

16.1 Introduction

An oscillator is an important ingredient in communications systems. It is used to modulate and de-modulate information onto an appropriate carrier and it is used in all timing aspects of the communication link. An oscillator is an *autonomous* circuit that converts DC power into a periodic waveform, in other words it generates an output waveform when only DC power is applied. We will initially restrict our attention to a class of oscillators that generate a sinusoidal waveform. The period of oscillation is determined by a high-Q *LC* tank or a resonator (crystal, cavity, T-line, etc.). An oscillator is characterized by its oscillation amplitude (or power), frequency, the frequency “stability” versus temperature and from part-to-part, phase noise, and sensitivity to supply variations (Fig. 16.1). Most of these terms will be defined later. A common class of oscillators is the so-called Voltage Controlled Oscillators, commonly known as VCOs, are oscillators whose output frequency

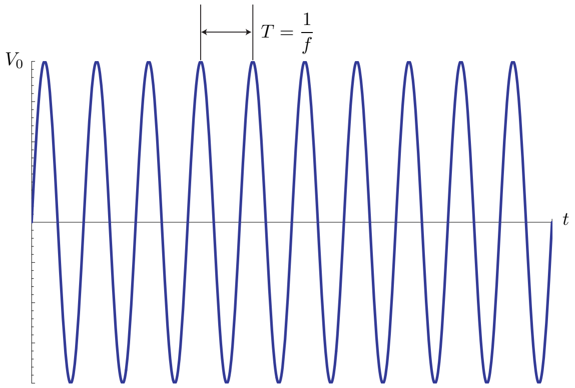


Figure 16.1: The output of an ideal harmonic oscillator is a sinusoidal waveform with a well defined amplitude and period of oscillation.

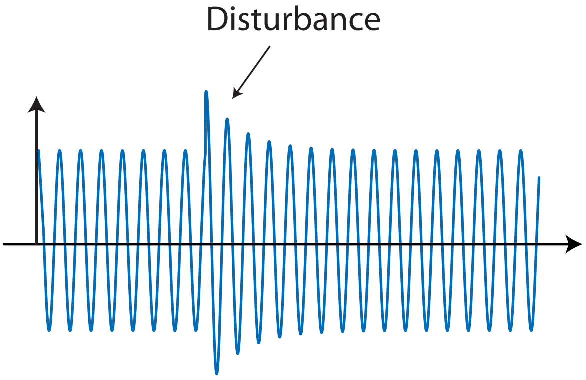


Figure 16.2: A stable oscillator is characterized by amplitude stability, or the ability to return to a fixed amplitude even if the face of disturbances.

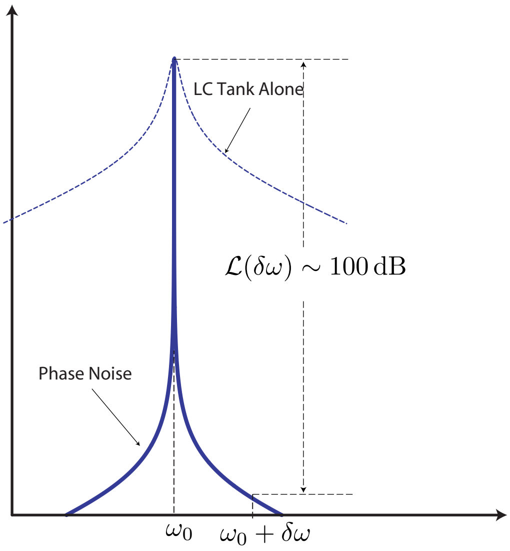


Figure 16.3: The spectrum of a practical oscillator deviates from a pure tone due to phase noise.

is tunable, usually by means of tuning a variable capacitor (varactor)[1] VCOs are also characterized by their tuning range, specified as a fraction of the fundamental frequency. Some applications require a spectrally pure output frequency, so the harmonic content of oscillator is of practical interest.

Generically, a good oscillator is stable in that its frequency and amplitude of oscillation do not vary appreciably with temperature, process, power supply, and external disturbances. As shown in Fig. 16.2, the amplitude of oscillation is particularly stable, always returning to the same value, even after a disturbance is injected into the tank. The reason for this amplitude stability will be elucidated in this chapter.

16.1.1 Phase Noise

Due to noise, a real oscillator does not have a purely periodic waveform, in other words the phase does not accumulate linearly with time. These slight deviations are known as phase noise. From a

[1] ICOs, or Current Controlled Oscillators, are also possible but are much less common. This is related to the fact that most varactors including pn-junctions or MOS-varactors use a voltage to control capacitance.

Figure 16.4: Jitter is the variation in the period of oscillation due to noise sources.

Figure 16.5: Jitter corrupts the eye diagram since the waveform never overlaps due to timing errors.

frequency domain perspect, the oscillator does not have a delta-function power spectrum, but rather a very sharp peak at the oscillation frequency. The power spectrum drops very quickly, though, as one moves away from the center frequency. E.g. a typical wireless communication oscillator (like a mobile phone) oscillator has a phase noise that is 100 dB down at an offset of only 0.01% from the carrier! It's important to realize that the spectral shape is not only related to the frequency response of an LC tank or crystal which is used to build the oscillator, but rather the spectral shape is determined by the sources of noise in the circuit and the positive feedback. This point will be revisited in Section **??**. Phase noise is often specified at a certain offset frequency from the carrier as the relative amplitude at the carrier compared to the carrier power. For example, if a 1 GHz carrier has 5 dBm of power and measured power at 1 MHz offset away is -100 dBm, then the phase noise is specfied as -105 dBc/$\sqrt{Hz}$ at 1 MHz offset. For reasons that will become clear later, the units of phase noise are dBc/$\sqrt{Hz}$.

16.1.2 Jitter

While phase noise characterizes the uncertainty in the frequency of oscillation, jitter characterizes the uncertainty of the period of oscillation. It's a completely equivalent characterization of phase noise but viewed from a different perspective. In Fig. 16.4 we see that the zero-crossing point of a real oscillator deviates from an ideal oscillator. If an eye diagram of this waveform is plotted as shown in Fig. 16.5, jitter causes the eye to close. Jitter is often specified in terms of the RMS value in the uncertainty of the period of oscillation. For instance a 1 GHz oscillator with 10ps RMS jitter means that the root mean squared value of the period will vary by 10ps/1ns = 1%.

16.2 Oscillator Viewed as a Linear Systems

Note that an LC tank alone is not a good oscillator. Due to loss, no matter how small, the amplitude of the oscillator decays (see Fig. 16.15). Even a very high Q oscillator can only sustain oscillations for about Q cycles[2]. For instance, an LC tank at 1 GHz has a $Q \sim 20$, can only sustain oscillations for about 20 ns. Even a resonator with high $Q \sim 10^6$, will only sustain oscillations for about 1 ms.

16.2.1 Feedback Perspective

Many oscillators can be viewed as feedback systems as shown in Fig. reffig:feedbackosc. The oscillation is sustained by feeding back a fraction of the output signal, using an amplifier to gain the signal, and then injecting the energy back into the tank. The transistor "pushes" the LC tank with just about enough energy to compensate for the loss. Note that this is a positive feedback system.

16.2.2 Negative Resistance Perspective

Another perspective is to view the active device as a negative resistance generator. In steady state, the losses in the tank due to conductance G are balanced by the power drawn from the active device through the negative conductance $-G$, as shown in Fig. 16.8.

[2] Recall that Q is related to the ratio of average energy stored in the tank versus the loss per cycle

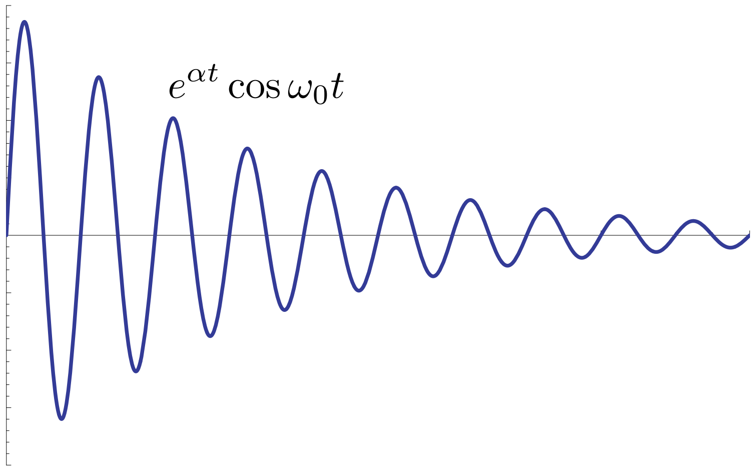


Figure 16.6: Any linear system with loss cannot sustain a constant amplitude oscillation.

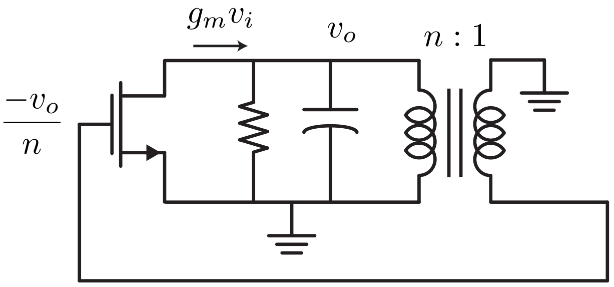


Figure 16.7: An oscillator is viewed as a feedback system, where a fraction of the oscillator output signal is fed back, amplified, and then added in phase with the original signal.

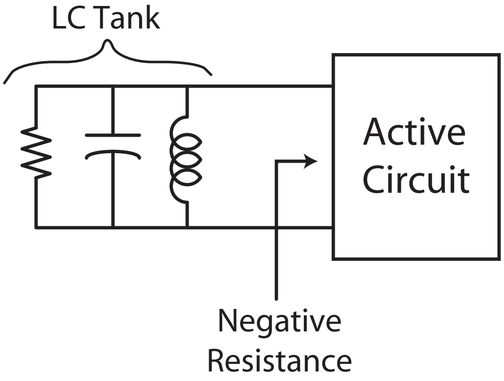


Figure 16.8: An oscillator can be viewed as a lossy *LC* tank connected to a generator that presents negative conductance to compensate for the losses in the *LCR* tank.

Figure 16.9: (a) A diode-connected transistor is equivalent to a two-termianl conductance g_m. (b) If the gate (base) voltage is inverted, the transistor presents a negative conductance $-g_m$.

Figure 16.10: Negative resistance realized by inverting the gate signal through a (a) transformer or (b) second common source amplifier. (c) Instead of inverting the gate, we can hold it constant and drive the source (emitter) terminal.

The concept of negative conductance (or resistance) may seem strange and unphysical at first, but we will find it to be a very useful way to analyze a certain class of oscillators. First note that when you apply a voltage to a negative conductor, current flows *out* of the device, providing energy to the source. Thus such a device is an *active* rather than *passive* device. How can one realize a negative conductor? One simple technique is to use a transconductance (such as a transistor) in a modified “diode" configuration. As shown in Fig. 16.9a, a diode connected FET has an output conductance of g_m since the gate voltage and output voltage are driven as one terminal. If we could simply invert the polarity of the gate voltage, as shown in Fig. 16.9b, then the output conductance becomes negative, or $-g_m$. A transformer or a common source amplifier does this job nicely, as shown in Fig. 16.10a-b. If there is voltage gain through the inversion process, then the conductance increases to $-g_m \times n$ (n is the voltage gain). Alternatively, instead of moving the gate voltage out of phase with the drain voltage, we can achieve the same effect if we move the source voltage in phase with the drain voltage while holding the gate constant, as shown in Fig. 16.10c, where a capacitive divider is used to couple the drain voltage to the source. This configuration has a voltage attenuation, so the output conductance is $-g_m/n$.

In the above circuits we see that we are effectively using feedback to realize a negative resistance. We are thus viewing the problem from the other direction, which shows the equivalence of the two concepts. The negative resistance approach, though, is particularly useful when the device under consideration is a two-port rather than a three-port active device. Examples include tunneling diodes where a negative *incremental* resistance is obtained due to a negative slope in the *I*-*V* region.

16.2.3 Wave Reflection Perspective

If we view an *LCR* tank as a scattering one-port element, then we can say that due to loss, the reflection coefficient is always less than unity

$$|\Gamma_L(f)| < 1 \tag{16.1}$$

If this impedance is connected to an element with a reflection coeffient larger than unity $|\Gamma_G(f)| > 1$ such that

$$|\Gamma_L(f)\Gamma_G(f)| > 1 \tag{16.2}$$

then a waveform disturbance traveling back and forth between these elements will regenerate and grow in amplitude, shown in Fig. **??**. In fact, if the reflected wave is in phase with the incoming wave, then they will grow the most. If at a particular frequency $\Gamma_L(f_0)$ is real (resonance), then $\Gamma_G(f_0)$ should also be real. Note that a real reflection coefficient with magnitude larger than unity corresponds to a negative resistance

$$Z_G = Z_0 \frac{1+\Gamma_G}{1-\Gamma_G} = -Z_0 \frac{1+\Gamma_G}{|\Gamma_G - 1|} \tag{16.3}$$

which shows that all three perspectives are in fact equivalent.

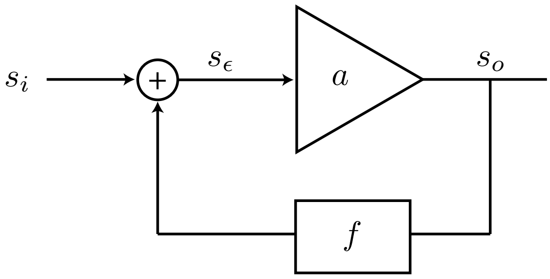


Figure 16.11: An ideal feedback system.

Example 1: Feedback Example

Consider an ideal feedback system with forward gain $a(s)$ and feedback factor $f(s)$ as shown in Fig. 16.11. The closed-loop transfer function is given by

$$H(s) = \frac{a(s)}{1+a(s)f(s)} \tag{16.4}$$

Suppose the forward gain transfer function is composed of three identical real negative poles with magnitude $|\omega_p| = 1/\tau$ and a frequency independent feedback factor f

$$a(s) = \frac{a_0}{(1+s\tau)^3} \tag{16.5}$$

Deriving the closed-loop gain, we have

$$H(s) = \frac{a_0}{(+s\tau)^3 + a_0 f} = \frac{K_1}{(1-s/s_1)(1-s/s_2)(1-s/s_3)} \tag{16.6}$$

where $s_{1,2,3}$ are the poles of the feedback amplifier.

Solving for the poles

$$(1+s\tau)^3 = -a_0 f \tag{16.7}$$

$$1+s\tau = (-a_0 f)^{\frac{1}{3}} = (a_0 f)^{\frac{1}{3}}(-1)^{\frac{1}{3}} \tag{16.8}$$

$$(-1)^{\frac{1}{3}} = -1,\ e^{j60^\circ},\ e^{-j60^\circ} \tag{16.9}$$

The poles are therefore

$$s_1,\ s_2,\ s_3 = \frac{-1-(a_0 f)^{\frac{1}{3}}}{\tau},\ \frac{-1+(a_0 f)^{\frac{1}{3}} e^{\pm j60^\circ}}{\tau} \tag{16.10}$$

If we plot the poles on the s-plane as a function of the DC loop gain $T_0 = a_0 f$, we generate a *root locus*. For $a_0 f = 8$, the poles are on the $j\omega$-axis with value

$$s_1 = -3/\tau \tag{16.11}$$

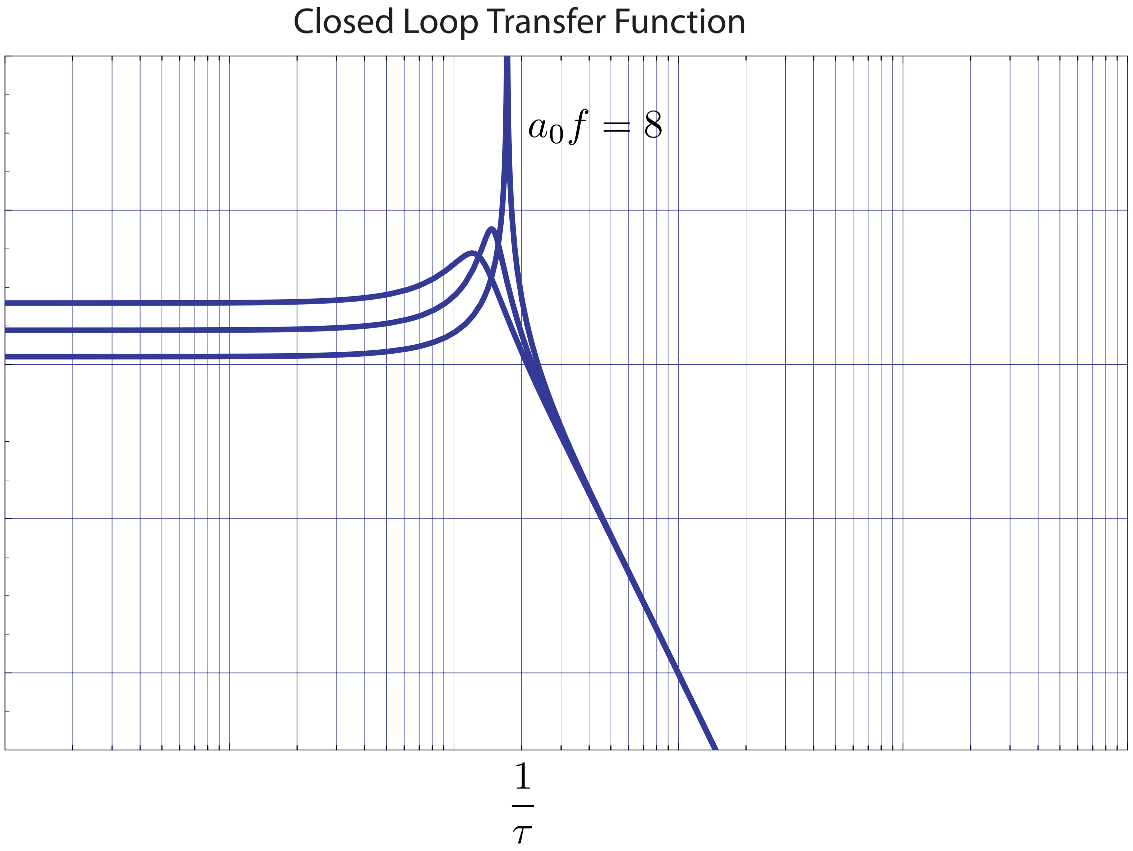


Figure 16.12: The magnitude of the closed loop transfer function as a function of loop gain.

$$s_{2,3} = \pm j\sqrt{3}/\tau \tag{16.12}$$

For $a_0 f > 8$, the poles move into the right-half plane (RHP).

In the frequency domain, we see that our example feedback amplifier has a transfer function

$$H(j\omega) = \frac{a(j\omega)}{1 + a(j\omega)f} \tag{16.13}$$

If the loop gain $a_0 f = 8$, then we have with purely imaginary poles at a frequency $\omega_x = \sqrt{3}/\tau$ where the transfer function $a(j\omega_x)f = -1$ blows up. Apparently, the feedback amplifier has infinite gain at this frequency.

16.2.4 Natural Response

When a given a transfer function

$$H(s) = \frac{K}{(s - s_1)(s - s_2)(s - s_3)\cdots} = \frac{a_1}{s - s_1} + \frac{a_2}{s - s_2} + \frac{a_3}{s - s_3} + \cdots \tag{16.14}$$

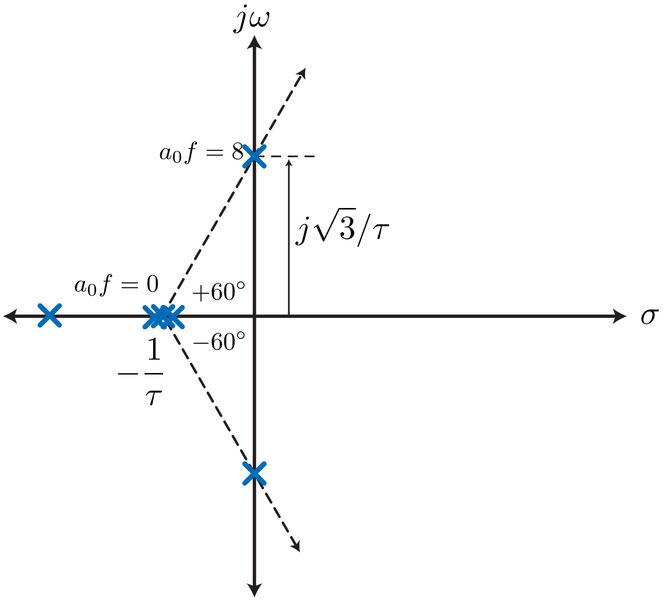


Figure 16.13: The root locus of the system as a function of loop gain.

the total response of the system can be partitioned into the *natural response* and the forced response

$$s_0(t) = f_1(a_1 e^{s_1 t} + a_2 e^{s_2 t} + a_3 e^{s_3 t} + \cdots) + f_2(s_i(t)) \tag{16.15}$$

where $f_2(s_i(t))$ is the forced response whereas the first term $f_1()$ is the natural response of the system, even in the absence of the input signal. The natural response is determined by the initial conditions of the system.

Real LHP Poles

Stable systems have all poles in the left-half plane (LHP). Consider the natural response when the pole is on the negative real axis, such as s_1 for our examples. The response is a decaying exponential that dies away with a time-constant determined by the pole magnitude (Fig. 16.14).

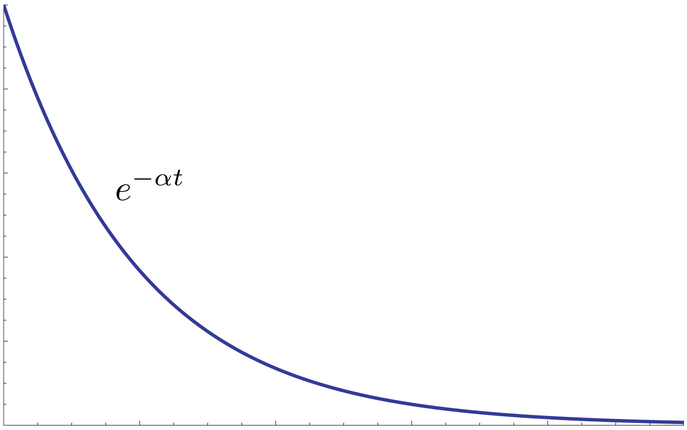


Figure 16.14: A real axis left-hand plane pole gives rise to a decaying exponential natural response.

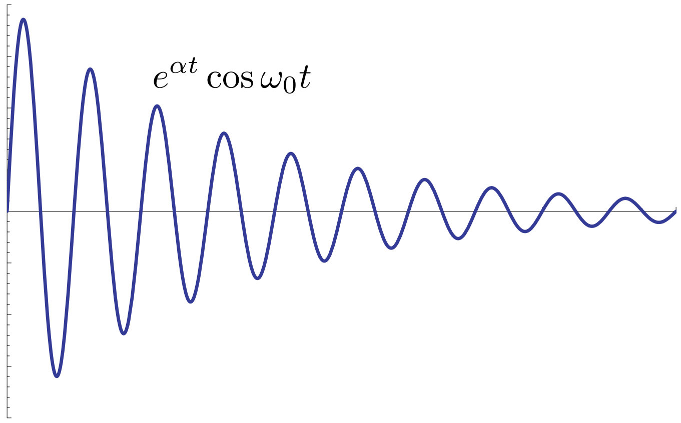


Figure 16.15: A pair of left-hand plane complex conjugate poles correspond to a decaying oscillatory natural response.

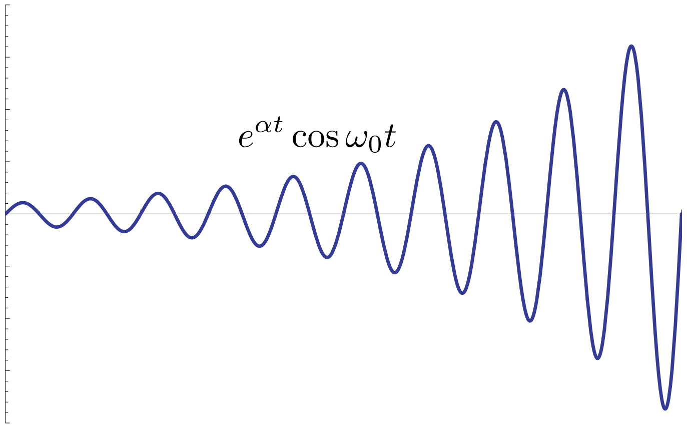


Figure 16.16: Complex conjugate poles in the right-half plane result in a growing oscillatory natural response.

Complex Conjugate LHP Poles

Suppose $s_{2,3}$ are a complex conjugate pair

$$s_2,\ s_3 = \sigma \pm j\omega_0 \tag{16.16}$$

We can group these responses since $a_3 = \overline{a_2}$ into a single term

$$a_2 e^{s_2 t} + a_3 e^{s_3 t} = K_a e^{\sigma t}\cos\omega_0 t \tag{16.17}$$

When the real part of the complex conjugate pair σ is negative, the response also decays exponentially (Fig. 16.15).

Complex Conjugate Poles (RHP)

When σ is positive (RHP), the natural response is an exponential growing oscillation at a frequency determined by the imaginary part ω_0. For the example amplifier with three identical poles, if feedback is applied with loop gain $T_0 = a_0 f > 8$, the amplifier will oscillate.

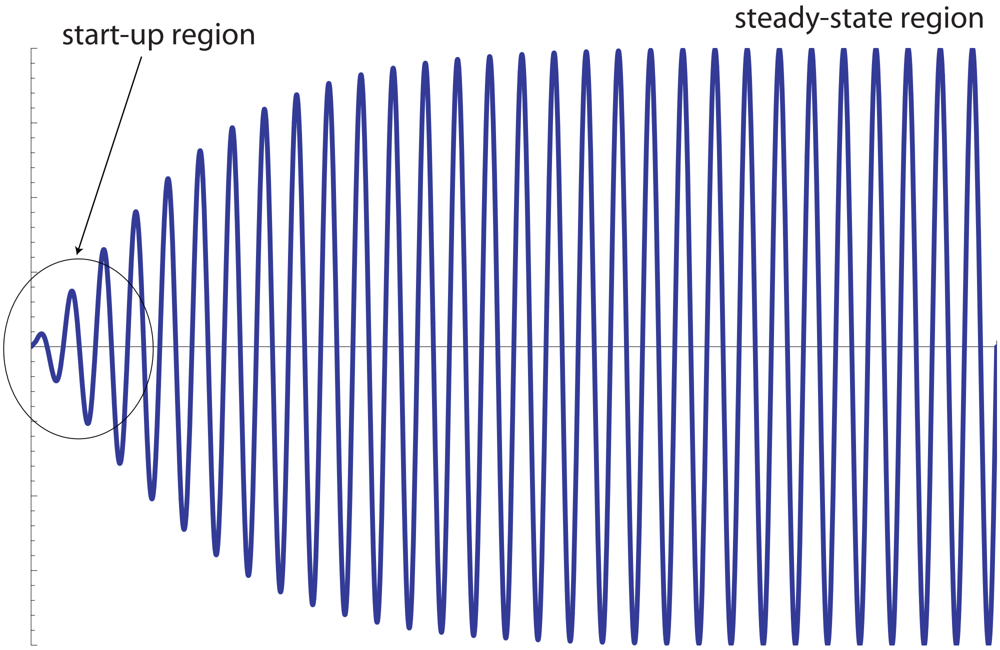


Figure 16.17: Any real oscillator will limit at a certain oscilllation amplitude which cannot be predicted based on linear analysis.

16.2.5 Oscillation Build Up: Failure of Linear Analysis

In a real oscillator, the amplitude of oscillation initially grows exponentially as our linear system theory predicts. This is expected since the oscillator amplitude is initially very small and such theory is applicable. But as the oscillations become more vigorous, the non-linearity of the system comes into play, and amplitude limiting, as shown in Fig. 16.17, comes into play. We will analyze the steady-state behavior, where the system is non-linear but periodically time-varying.

16.2.6 An Extended Example: Transistor LC Oscillator

Consider the transistor *LC* oscillator shown in Fig. 16.18. The emitter resistor is bypassed by a large capacitor at AC frequencies. The base of the transistor is conveniently biased through the transformer windings. The LC oscillator uses a transformer for feedback. Since the amplifier gain has a phase shift of 180°, the feedback transformer needs to provide an additional phase shift of 180° to provide positive feedback. The AC equivalent circuit is shown in Fig. 16.19.

At the oscillation frequency, the AC equivalent circuit can be further simplified as shown in Fig. 16.20. The transformer winding inductance L resonates with the total capacitance in the circuit. The resistor R_T is the equivalent tank impedance at resonance.

To analyze the transfer function, we use the small-signal equivalent circuit shown in Fig. 16.21. The forward gain is given by $a(s) = -g_m Z_T(s)$, where the tank impedance Z_T includes the loading effects from the input of the transistor

$$R = R_0 || R_L || n^2 R_i \tag{16.18}$$

$$C = C_L + \frac{C_i}{n^2} \tag{16.19}$$

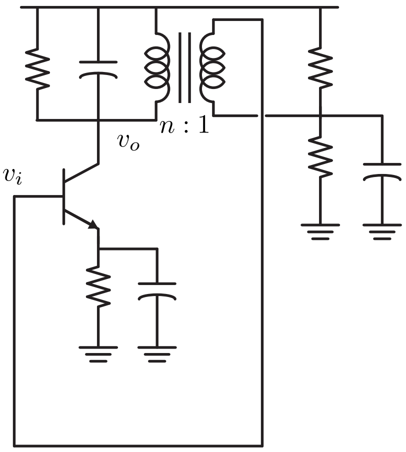


Figure 16.18: A transformer based *LC* oscillator.

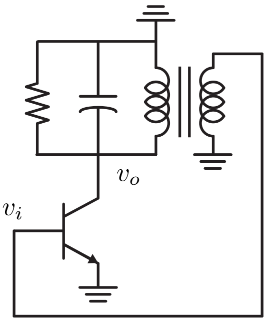


Figure 16.19: AC equivalent circuit of *LC* oscillator .

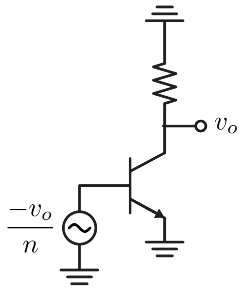


Figure 16.20: AC equivalent circuit at resonance.

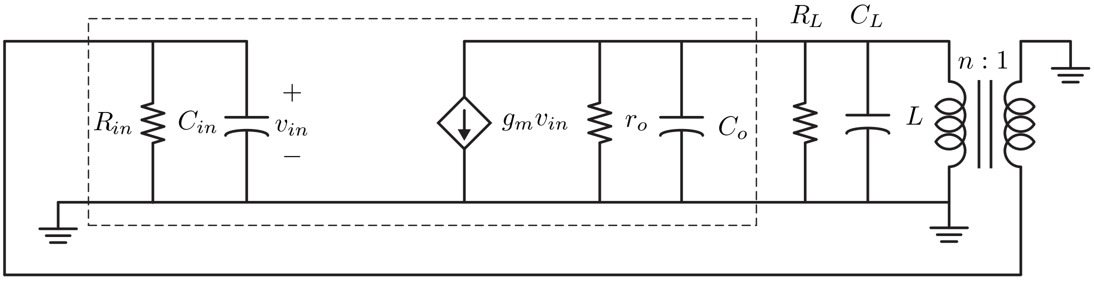


Figure 16.21: Complete small-signal model of transformer based *LC* oscillator.

The tank impedance is therefore

$$Z_T(s) = \frac{1}{sC + \frac{1}{R} + \frac{1}{Ls}} = \frac{Ls}{1 + s^2LC + sL/R} \tag{16.20}$$

The loop gain is given by

$$af(s) = \frac{-g_m R}{n} \frac{\frac{L}{R}s}{1 + \frac{L}{R}s + s^2LC} \tag{16.21}$$

The loop gain at resonance is the same as the DC loop gain

$$A_\ell = \frac{-g_m R}{n} \tag{16.22}$$

The closed-loop transfer function is given by

$$H(s) = \frac{-g_m R \frac{L}{R}s}{1 + s^2LC + s\frac{L}{R}(1 - \frac{g_m R}{n})} \tag{16.23}$$

The denominator can be written as a function of A_ℓ

$$H(s) = \frac{-g_m R \frac{L}{R}s}{1 + s^2LC + s\frac{L}{R}(1 - A_\ell)} \tag{16.24}$$

Note that as $n \to \infty$, the feedback loop is broken and we have a tuned amplifier. The pole locations are determined by the tank Q. The closed loop gain is plotted in Fig. 16.22 as a function of frequency and with different values of A_ℓ.

If $A_\ell = 1$, then the denominator loss term cancels out and we have two complex conjugate imaginary axis poles

$$1 + s^2LC = (1 + sj\sqrt{LC})(1 - sj\sqrt{LC}) \tag{16.25}$$

For a second-order transfer function, notice that the magnitude of the poles is constant, so they lie on a circle in the s-plane

$$s_1, s_2 = \frac{-a}{2b} \pm \frac{a}{2b}\sqrt{1 - \frac{4b}{a^2}} = \frac{-a}{2b} \pm j\frac{a}{2b}\sqrt{\frac{4b}{a^2} - 1} \tag{16.26}$$

$$|s_{1,2}| = \sqrt{\frac{a^2}{4b^2} + \frac{a^2}{4b^2}(\frac{4b}{a^2} + 1)} = \sqrt{\frac{1}{b}} = \omega_0 \tag{16.27}$$

which facilitates the root locus diagram shown in Fig. 16.23.

We see that for $A_\ell = 0$, the poles are determined by the tank Q and lie in the LHP. As A_ℓ is increased, the action of the positive feedback is to boost the gain of the amplifier and to decrease the bandwidth. Eventually, as $A_\ell = 1$, the loop gain becomes infinite in magnitude.

16.3 Oscillator Steady State Analysis

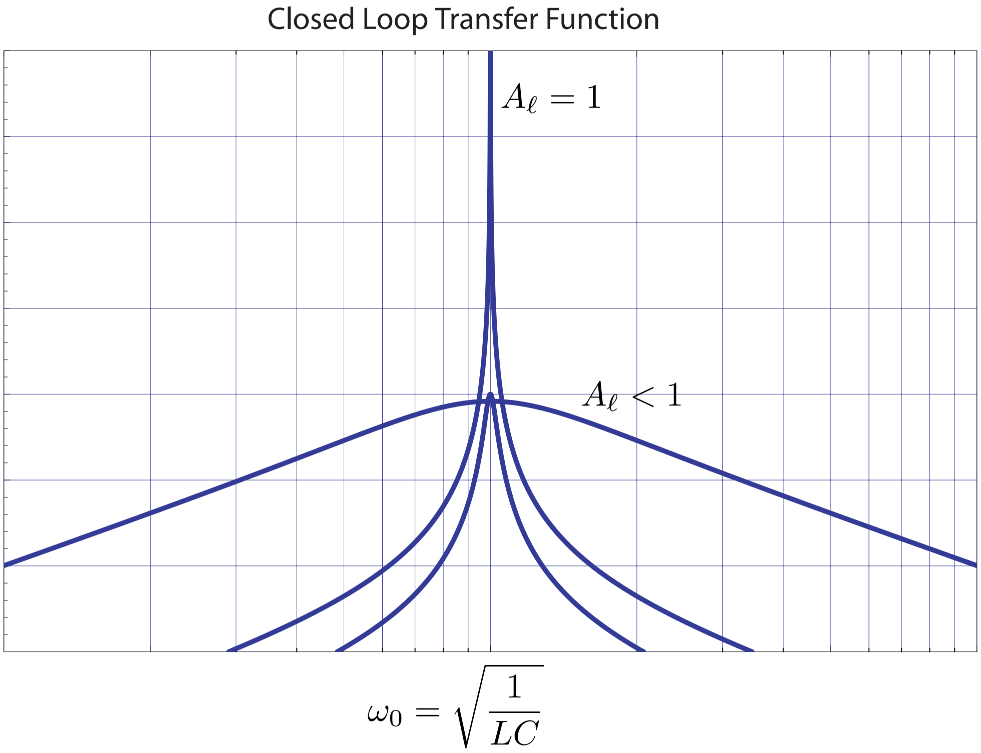


Figure 16.22: The close loop transfer function as a function of frequency and loop gain A_ℓ.

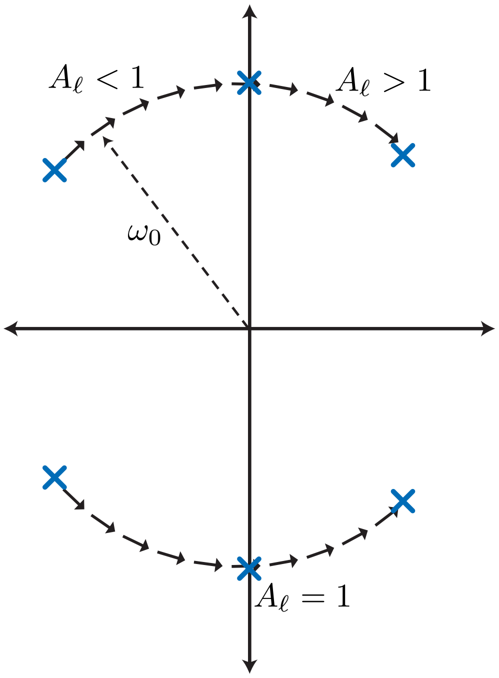
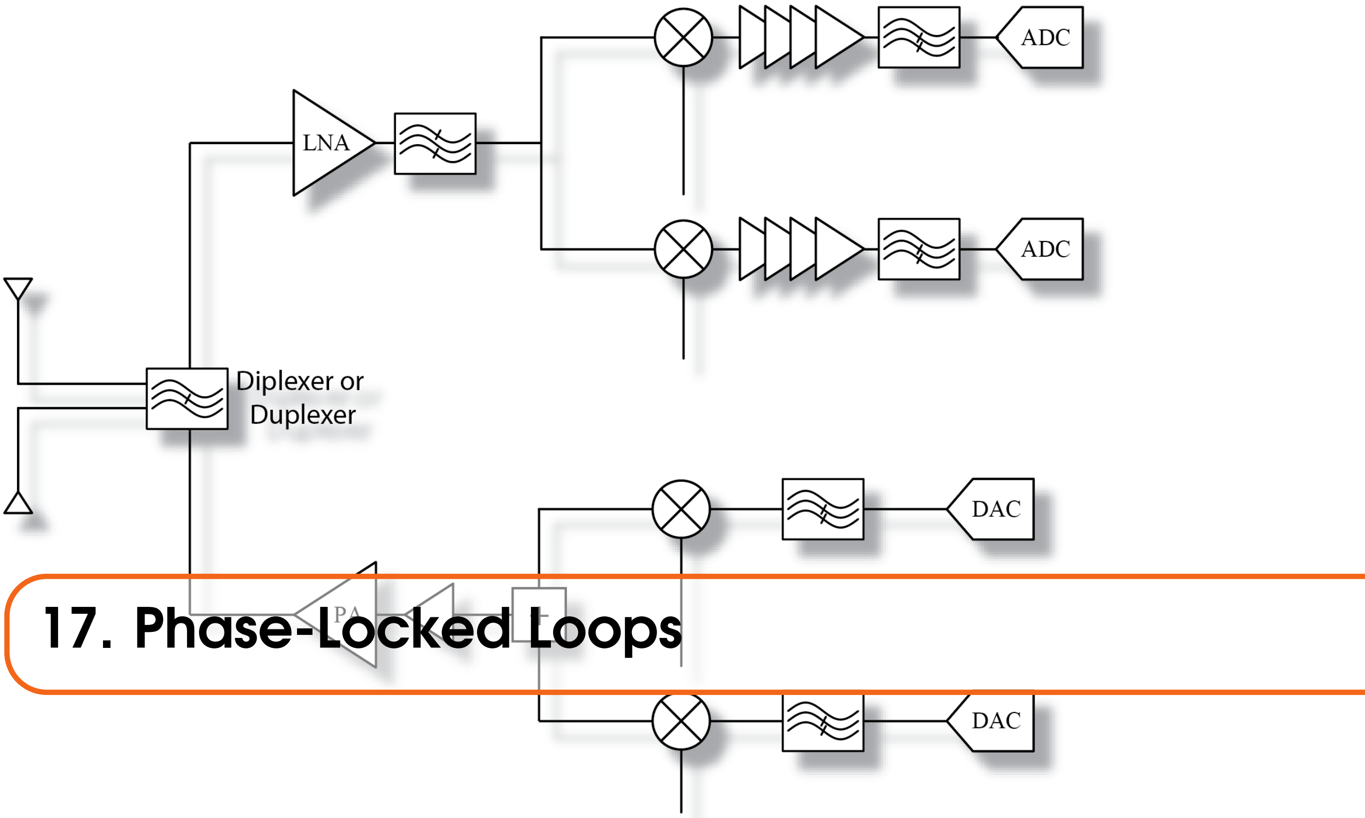


Figure 16.23: Root locus for a second-order transfer function analyzed in connection with the *LC*.



17. Phase-Locked Loops

17.1 Introduction

Phase Locked Loops (PLL) are ubiquitous circuits used in countless communication and engineering applications. As shown in the block diagram of Fig. 17.1, the key components include a voltage controlled oscillator (VCO), a frequency divider, a phase detector (PD), and a loop filter. A PLL is a truly mixed-signal circuit, involving the co-design of RF, digital, and analog building blocks. This chapter will explain all of these building blocks in detail but for now you can think of the PLL as a negative feedback system that aligns the phase of the VCO divided clock with a reference signal. The frequency divider simply generates one rising edge for *N* edges of the VCO, and it's usually implemented as a sigital circuit like a counter. The phase detector generates an error signal proportional to the phase difference between the rising edge of the divided clock and the reference clock and the control voltage of the VCO increases or decreases the frequency of the VCO to align the phase. Notice that we don't directly control the phase of the VCO but its frequency, and since frequency is the derivative of phase, the VCO is an integrator in the loop.

There are many applications that utilize a PLL as their core component. Examples include

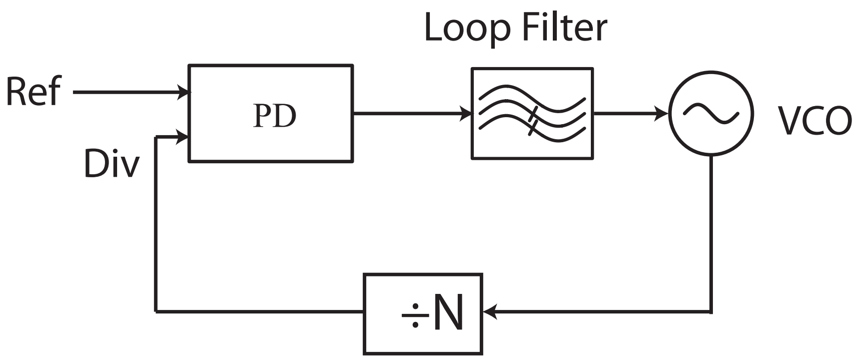


Figure 17.1: Block diagram of a generic PLL.

generating a clean, tunable, and stable reference (LO) frequency, a process referred to as *frequency synthesis*. The reference frequency is derived from a stable resonator, such as a crystal (XTAL) oscillator (XO). The VCO frequency varies with temperature and is process and voltage dependent. By locking the frequency of the VCO to the XTAL, we can generate another (usually higher) frequency from frequency reference source.

Other applications include frequency modulation and demodulation. A PLL is a natural frequency modulator/de-modulator since the output frequency can be modulated linearly by changing the division ratio, which is usually a programmable digital divider. Other important applications include clock recovery for high speed communication, and the generation of phase synchronous clock signals in microprocessors. PLL's are even used in optical and mechanical systems, and even combinations such as electro-optical PLLs.

17.2 Frequency Synthesizer

In a frequency synthesizer, the VCO is usually realized using an *LC* tank (to get the best phase noise), or alternatively a ring oscillator (higher phase noise, smaller area). The reference is derived from a precision XTAL oscillator (see Sect. **??**). The divider brings down the high frequency of the VCO signal to the range of the reference frequency. The PD compares the phase and produces an error signal, which is smoothed out by the loop filter and applied to the VCO. When the system locks, the output phase of the VCO is locked to the XTAL, which in turn also means that the frequency is also locked to a multiple of the input reference frequency. In other words, the output frequency f_{out} is an integer multiple of the reference f_{ref} since in lock we have:

$$f_{ref} = f_{out}/N \tag{17.1}$$

or

$$f_{out} = N \times f_{ref} \tag{17.2}$$

Since the division ratio N is a digital circuit, we can vary the output frequency by changing the division ratio N.

17.2.1 Programmable Divider

By making the divider N programmable, we can tune the VCO frequency in either integer steps of the reference (integer-N architecture) or in fractional amounts (fractional-N architecture).

$$\Delta f = (N+p)f_{ref} - Nf_{ref} = pf_{ref} \tag{17.3}$$

In a fractional divider, $p < 1$ and is realized by dithering the divider between N and $N+1$ using a sigma-delta modulator. In practice, the programmable divider is made of up asynchronous high-speed dividers followed by programmable CMOS dividers (counters). The high speed dividers are sometimes in CML, which runs faster than CMOS, and has superior noise immunity and generation due to the differential nature. Injection locked or TSPC dividers are also useful for very low power high frequency operation.

17.3 Linear PLL Model

A PLL is described by several parameters, such as the *locking range*, or the range of frequencies for which it will stay locked. The *capture range* is the frequency range for which it will lock from an initially unlocked state. The capture range is smaller than the locking range. These parameters are hard to derive analytically and require simulation. But the dynamics of the loop, such as settling

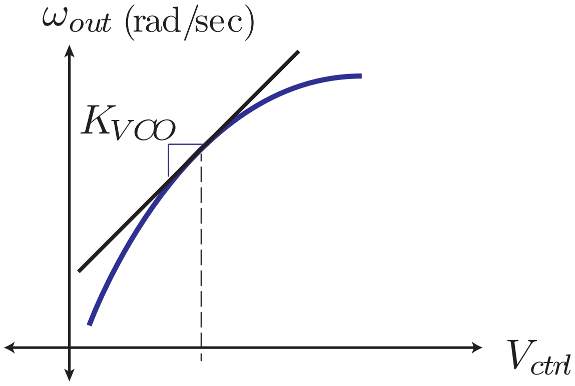


Figure 17.2: The VCO tuning curve is the output frequency versus control voltage.

time, and the noise transfer characteristics (phase noise), can be derived from a linear model. Therefore it is useful to derive a linear model by assuming the system is close to lock, or in lock. The most convenient variable is *phase*, and not frequency, in the linear model. Since phase and frequency are related, it's easy to go back and forth.

17.3.1 The VCO Linear Model (K_{VCO})

The VCO tuning curve is generally non-linear and given by a plot of output frequency versus control voltage (see Fig. 17.2). But when the PLL is in lock, the control voltage V_{ctrl} varies only around a small region around the lock point. We can therefore model the VCO linearly.

$$F_{VCO} = K_{VCO} V_{ctrl} \tag{17.4}$$

where

$$K_{VCO} = \frac{\partial F_{VCO}}{\partial V_{ctrl}} \tag{17.5}$$

Since we are interested in the phase, and observing that frequency is the time derivative of phase, we can derive an s-domain model as follows:

$$\Phi_{VCO} = \frac{1}{s} F_{VCO} = \frac{K_{VCO}}{s} V_{ctrl} \tag{17.6}$$

The VCO is therefore an implicit *integrator* in the loop. This is an important fact to consider when designing a PLL.

17.3.2 Divider Linear Model

From a voltage input-output characteristic, the divider is a non-linear block that simply acts like a counter. For N input edges, only one output edge occurs. But in terms of phase, it's a linear block

$$F_{Div} = \frac{F_{VCO}}{N} \tag{17.7}$$

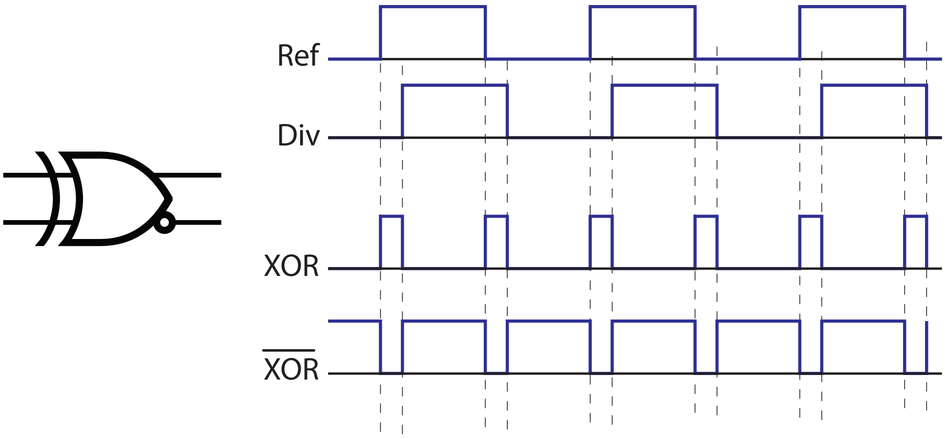
converting from frequency phase:

$$\Phi_{Div} = \int_{-\infty}^{t} F_{Div}(\tau) d\tau = \int_{-\infty}^{t} \frac{F_{VCO}}{N}(\tau) d\tau \tag{17.8}$$

or

$$= \frac{1}{N} \int_{-\infty}^{t} F_{VCO}(\tau) d\tau = \frac{1}{N} \Phi_{VCO} \tag{17.9}$$

This shows that the linear gain is just the division ratio.



(a) (b)

Figure 17.3: (a) A digital differential XOR circuit. (b) The waveforms of the XOR for two hypothetical input signals.

17.3.3 Multiplier Phase Detector

The most classic phase detector (PD) is a multiplier. Consider the product of two sinusoids offset by some phase ϕ. The product is simply given by

$$e(t) = AB\cos(\omega t)\cos(\omega t+\phi) = \frac{AB}{2}(\cos(\phi)-\cos(2\omega t+\phi)) \tag{17.10}$$

After a low-pass filter (LPF), the high frequency term at twice the frequency is filtered out, so we have

$$< e(t) > \approx \frac{AB}{2}\cos(\phi) \tag{17.11}$$

The slope of the phase detector around zero is given by

$$K_{PD} = \frac{de(t)}{d\phi} = -\frac{AB}{2}\sin\phi \tag{17.12}$$

In locked condition, the phase deviations are small and we can make the simple linear approximation

$$K_{PD} \approx -\frac{AB}{2} \tag{17.13}$$

Note that this system will lock the VCO onto the quadrature of the reference signal. The negative sign is not much concern, because it can be absorbed into other gain blocks which have positive or negative gains, depending on how they are designed. We must ensure that the overall loop has negative phase shift to form negative feedback. Some designers reserve the option to swap the inputs of the PD just to be sure they can change things in case they make an error!

17.3.4 XOR Phase Detector

The differential XOR gate acts very much like a multiplier. The best way to derive the transfer function is just to draw some ideal digital signals at the inputs and outputs and to find the average level of the output signal (see Fig. **??**). Since the XOR circuit generates an output if either input is one while the other is zero, it generates an output signal at a rate of twice the input frequency.

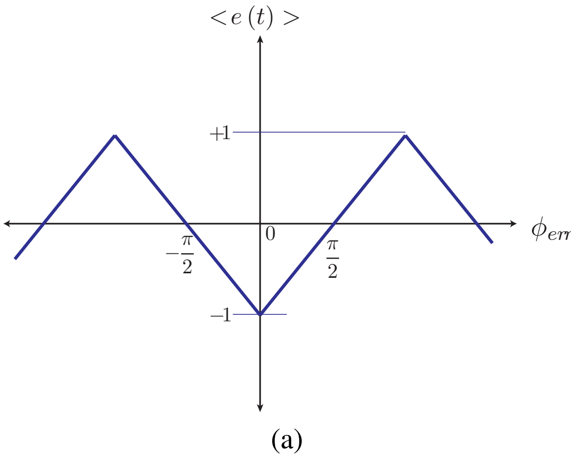


Figure 17.4: (a) The transfer function of the XOR phase detector. (b) Sample waveforms at the input and output of the XOR PD.

There's also a DC shift, which depends on the relative balance of the waveforms. Most importantly, the average value of the output is a linear function of the phase difference, which is exactly what we want.

A sketch of the transfer curve shows that the system will also lock in quadrature (if perfectly balanced). When the signals are in phase, the output of the XOR is always zero while $\overline{XOR}$ is always one. On the other hand, if the signals are 180° output of phase, then the opposite is true. In the middle, when the signals are in quadrature, the output of the PD is zero. That's because the difference $XOR - \overline{XOR}$ becomes a balanced signal (50% duty cycle) which has an average of zero.

The slope of the line determines the PD gain, and since the output varies from 1 to -1 when the input varies of π radians, we have: $K_{PD} = \frac{1}{\pi}$ As noted earlier, this system locks in quadrature as the duty cycle of the positive and negative outputs is balanced, which produces a zero average output. Also note that the PD function is also periodic, much like the multiplier. In fact, the schematic of a XOR (CML) and a multiplier are very similar except for the signal levels.

As we have seen, the phase detector is actually a non-linear block that only extracts the phase on an average sense. We use the PD average behavior in the linear model. On average, the PD produces an error signal by taking the difference between the reference phase and the divided VCO phase. The PD gain is related to the slope of the transfer function

$$e(t) = K_{PD}(\Phi_{ref}(t) - \Phi_{Div}) \tag{17.14}$$

17.3.5 Loop Filter

The loop filter $H(s)$ is an linear filter that smooths out the error signal and is a critical part of the system under the control of the designer. Passive RC and active filters are both used to realize the loop filter. Ideally the voltage on the control node of a VCO should settle to a DC value to avoid *reference spurs*. In other words, if we apply a periodic waveform on the control line, we get FM side-bands which are undesirable. Since the PD block is non-linear and non-ideal, even in lock it can produce a waveform that needs to be filtered to minimize reference spurs.

17.3.6 Complete Linear Phase Model

Given all the linear models we have derived, we can form the block diagram shown in Fig. 17.5. Note that this model is in the phase domain, rather than the voltage or current domain. In other words, this is an abstract model that describes how the phase of the system evolves, rather than the

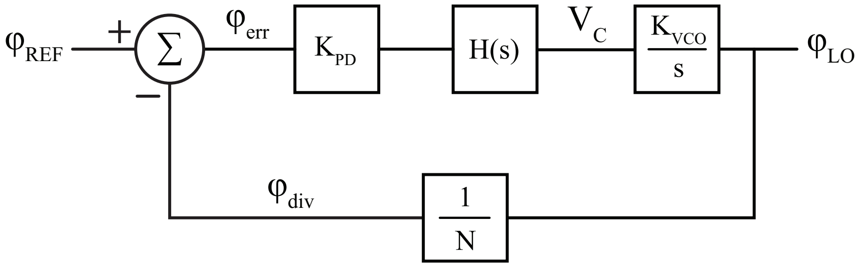


Figure 17.5: The complete block diagram of the phase domain PLL linear model.

voltage or current at any particular point in the circuit. The loop gain is given by

$$A(s) = \frac{K_{PD}H(s)K_{VCO}}{Ns} \tag{17.15}$$

It is worth mentioning that while the filter is not in the phase domain, the combination of the filter and VCO are effectively phase domain. In other words, a phase error voltage at the input of the filter generates a voltage that is applied to the control knob of the VCO, which in turn generates an output phase that is an integral of the voltage applied. This point is important to understand because it means the output of the phase detector should generate an error voltage. Of course it's possible to use current mode or combinations of current/voltage mode circuits, but the overall output should be interpreted correctly in the phase domain.

17.3.7 Closed-Loop Gain

The closed loop gain is given by

$$G(s) = \frac{A}{1+Af} = \frac{\frac{K_{PD}H(s)K_{VCO}}{s}}{1+\frac{K_{PD}H(s)K_{VCO}}{Ns}} \tag{17.16}$$

This is simplified to

$$G(s)/N = \frac{K_{PD}H(s)\frac{K_{VCO}}{N}}{s+K_{PD}H(s)\frac{K_{VCO}}{N}} \tag{17.17}$$

From our knowledge of linear systems, we know that all the properties of the system are a function of the poles and zeros of the transfer function. In practice we will design our components K_{VCO} and the loop filter ($H(s)$) to have the desired response.

Also note that the transfer function should ideally be such that the output phase tracks the input phase (from the reference). That's the entire point of the PLL, which means the steady-state behavior of the system should be carefully examined. The steady-state response of the system is given by the Final Value Theorem. For instance, the final value of the error signal is given by

$$\lim_{s\to 0} s\cdot G(s) = N \tag{17.18}$$

or

$$\lim_{s\to 0} s\cdot \frac{K_{PD}H(s)\frac{K_{VCO}}{N}}{s+K_{PD}H(s)\frac{K_{VCO}}{N}} = 1 \tag{17.19}$$

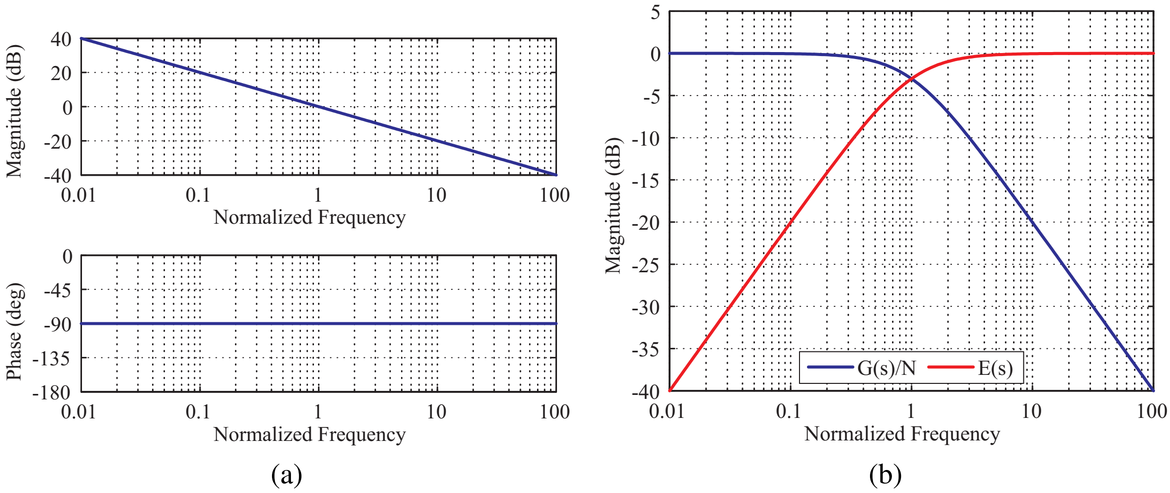


Figure 17.6: (a) The open-loop and (b) closed-loop input and error transfer functions for a PLL without a loop filter.

17.3.8 Noise Transfer Function

If we consider the phase noise coming out of the VCO, its transfer function to the output is different from the input and given by (also the transfer function for the error signal)

$$E(s) = \frac{1}{1+A(s)} = \frac{s}{s+K_{PD}H(s)\frac{K_{VCO}}{N}} \tag{17.20}$$

The VCO noise is therefore attenuated by the loop gain, which is very nice since the reference is usually more spectrally pure than the VCO (it is typically constructed using a high-Q quartz resonator) whereas the VCO uses a low Q on-chip tank (inductor + varactor). Note that when the loop gain drops (outside of the bandwidth of the PLL), the noise of the PLL is essentially governed by the free-running noise of the VCO.

17.4 Type I PLL

In this section we'll review PLL dynamics for a so-called type I PLL, a PLL consisting of a single integrator in the loop. As we'll demonstrate below, for a type I PLL the integrator is provided by the VCO itself.

17.4.1 Case 1: No Loop Filter

If we omit a loop filter, $H(s) = 1$, and the loop gain is given by $A(s) = \frac{1}{Ns}K_{PD}K_{VCO}$ and the closed loop gain and error function are low-pass and high-pass respectively

$$G(s) = \frac{K_{PD}K_{VCO}}{s+K_{PD}\frac{K_{VCO}}{N}} \tag{17.21}$$

$$E(s) = \frac{s}{s+K_{PD}\frac{K_{VCO}}{N}} \tag{17.22}$$

The plots of the open- and closed-loop transfer functions are shown in Fig. 17.6.

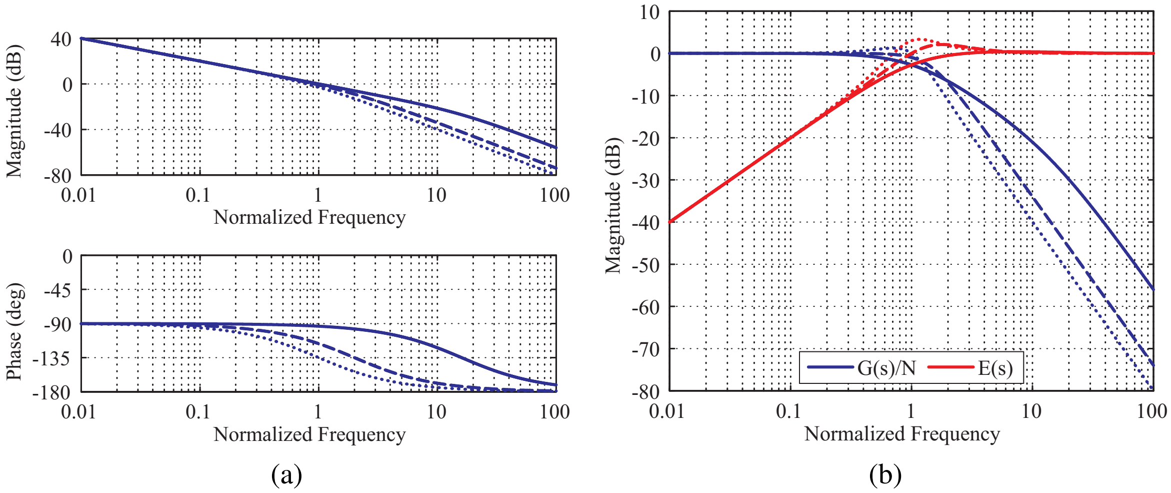


Figure 17.7: (a) The open-loop and (b) closed-loop transfer / error transfer functions for a PLL utilizing a single pole LPF $H(s)$.

The system has a 90° phase margin, and the loop bandwidth is given by

$$\omega_c = \frac{K_{PD}K_{VCO}}{N} \tag{17.23}$$

Within the loop bandwidth, the output phase follows the input phase and the noise of the VCO is rejected. Outside of the band, the phase is determined by the free running VCO. It's important to realize that the presence of the VCO makes the system low-pass in nature, because the VCO is essentially an integrator. The overall response is simple (neglecting higher order poles) and robust but the loop bandwidth, and hence the settling behavior, is fixed and cannot be changed once we design the VCO K_{VCO}. Not only does the bandwidth determine the dynamic settling behavior, but it also sets the bandwidth over which we reject other sources of noise in the system. So far we have only considered the VCO phase noise (see Sect. **??**), but every source of non-ideality is high-pass filtered in essentially the same manner and so the bandwidth of the loop is of paramount importance.

17.4.2 Case 2: Single Pole LPF

To gain more control over the loop bandwidth, let's see what happens if we design $H(s)$ to be a simple first order transfer function, or a simple single pole LPF:

$$H(s) = \frac{1}{1+\frac{s}{\omega_p}} \tag{17.24}$$

This renders the closed-loop response to be a second order function

$$G(s) = \frac{\omega_0^2}{s^2 + \frac{\omega_0 s}{Q} + \omega_0^2} \tag{17.25}$$

The natural frequency is given by

$$\omega_0 = \sqrt{K_{PD}\frac{K_{VCO}}{N}\omega_p} \tag{17.26}$$

The Quality factor is given by

$$Q = \sqrt{\frac{K_{PD}K_{VCO}}{N\omega_p}} \tag{17.27}$$

Since the transfer function is second order, the dynamics are well known (peaking behavior). One adjusts ω_p and the loop gain to set the phase margin. The loop gain increase (reduces) phase margin for a given ω_p. Obviously we don't want the Q to be too large, otherwise the system will *ring*. As the designer, we pick appropriate values for ω_0 and Q to meet our specifications. Keep in mind that these small-signal parameters vary over the lock range of the PLL, so we have to ensure proper behavior over the entire range. In practice K_{VCO} is the most important variation to take into account as most VCO's do not have a linear frequency/control voltage transfer function.

17.4.3 Steady-State Step Response

The steady-state response of the system is given by the Final Value Theorem. For instance, the final value of the error signal is given by

$$\lim_{t\to\infty} \Phi_e(t) = \lim_{s\to 0} sE(s)\phi_{in}(t) \tag{17.28}$$

If the input is a step function, $\phi_{in} = \Delta\phi/s$, so we have

$$\lim_{s\to 0} s \frac{s}{s + K_{PD}\frac{K_{VCO}}{N}H(s)} \frac{\Delta\phi}{s} \tag{17.29}$$

or

$$= \lim_{s\to 0} \frac{s}{s + K_{PD}\frac{K_{VCO}}{N}H(s)} \Delta\phi = 0 \tag{17.30}$$

17.4.4 Frequency Step

On the other hand, if the frequency of the input goes through a step change, that corresponds to a ramp function to the phase:

$$\phi_{in} = \frac{\Delta\omega}{s^2} \tag{17.31}$$

which means the at the error due to a frequency step is given by

$$\lim_{s\to 0} s \frac{s}{s + K_{PD}\frac{K_{VCO}}{N}H(s)} \frac{\Delta\omega}{s^2} = \frac{\Delta\omega}{K_{PD}\frac{K_{VCO}}{N}H(0)} \tag{17.32}$$

Unless the loop filter $H(s)$ has infinite DC gain, the loop will have a non-zero phase error if there is a frequency step. To remedy this, we should add another integrator into the loop. In fact, PLL's are characterized by the number of integrators in the loop. So far we have been using a type-I PLL, which has only 1 integrator (the VCO itself).

17.5 Type II PLL

As we demonstrated in the previous section, it is often desirable to have a PLL that can generate an output frequency that is precisely the same as an integer multiple of the reference. In order to tune the output frequency, we often simply change the division ration N, which is like applying a ramp perturbation to the phase of the system. In order to drive the phase error to zero, the loop needs a second integrator.

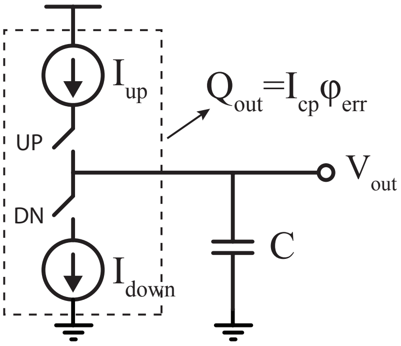


Figure 17.8: A simple integrator is realized using a current source and a capacitor. In practice, two current sources are used to generate both positive and negative charge flow onto the capacitor.

17.5.1 Charge Pump Integrator

A popular way to build a second integrator into the loop (note the first is the VCO) is to use a current source and a load capacitance, see Fig. 17.8:

$$V_{out} = V_C = \frac{Q_C}{C} = \frac{1}{C}\int_{-\infty}^{t} I(\tau)d\tau \tag{17.33}$$

In practice the current source is implemented by two sources that can pump current into and out of the capacitor. The *UP* and *DN* (Down) signals are controlled by a Phase Frequency Detector (PFD), described later. Essentially, an *UP* signal means the output should integrate up in voltage whereas a *DN* signal should do the opposite:

$$H(s) = I(s)\times\frac{1}{sC} = (I_{up} - I_{down})\frac{1}{sC} \tag{17.34}$$

where

$$I_{up} = \begin{cases} I_0 & UP > 0 \\ 0 & UP = 0 \end{cases} \tag{17.35}$$

where for simplicity we assume $UP > 0$ means the UP signal has arrived. Likewise:

$$I_{down} = \begin{cases} I_0 & DN > 0 \\ 0 & DN = 0 \end{cases} \tag{17.36}$$

Which means that the circuit is a tri-state circuit, either generating a positive current I_0 when $UP > 0$ and $DN = 0$, a negative current $-I_0$ when $DN > 0$, $UP = 0$, or zero output when both $UP > 0$ and $DN > 0$. A zero output is generated because the UP current source flows into the DN current source rather than into the load capacitor. Here we assume for simplicity that the currents sources are ideal and both generate the exact same current $I_{up} = I_{down} = I_0$ (no mismatch). Another way to generate a zero output is if both UP and DN are zero.

17.5.2 Lead/Lag Filter

Consider the lead-lag filter shown in Fig. **??**. The transfer function is from input current (provided by the charge pump) to the output voltage, so it's nothing but the impedance of the filter:

$$H(s) = \frac{1+\frac{s}{\omega_z}}{s(C_1+C_2)\left(1+\frac{s}{\omega_p}\right)} \tag{17.37}$$

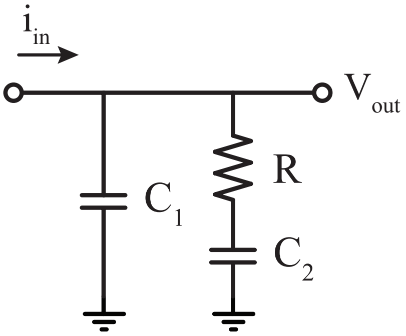


Figure 17.9: A simple RC lead-lag filter has two poles and a zero. One pole is at the origin, making this impedance an integrator.

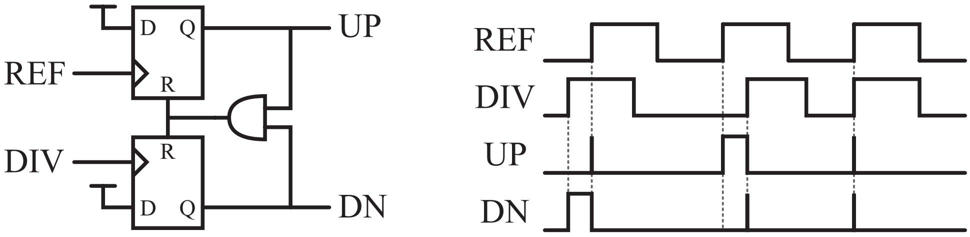


Figure 17.10: A realization of a phase-frequency detector (PFD) using flip-flops and an AND gate and the corresponding waveforms.

Clearly the transfer function has poles at the origin, due to the sum of the capacitors $C_1 + C_2$, which is desired to implement a second integrator in the loop. There's also a pole and a zero in the transfer function:

$$\omega_z = \frac{1}{R_1 C_2} \tag{17.38}$$

$$\omega_p = \frac{C_1 + C_2}{R_1 C_1 C_2} \tag{17.39}$$

By replacing the load capacitance of an integrator with this lead/lag filter, we can improve the stability of the loop. Typically the capacitor C_2 is much larger than C_1, so the pole occurs at a much higher frequency.

$$H(s) \approx \frac{1}{sC_2}(1 + s/\omega_z) \tag{17.40}$$

which is a useful simplification for our later analysis. It's the location of the zero ω_z, under our control, that we'll use to stabilize the loop.

17.5.3 Phase-Frequency Detector

As we discussed earlier, the charge pump is usually driven by a Phase-Frequency Detector (PFD), which is an edge sensitive circuit that measures the arrival time of the reference edge relative the divider edge. When a single edge arrives, the output goes high until the second edge arrives, at which time the output signal is reset to ground. The UP signal will output a one if the reference

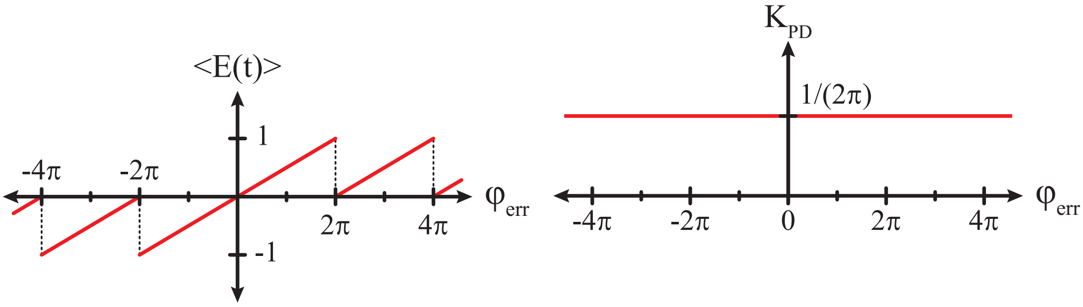


Figure 17.11: The average output of a PDF circuit measured by time averaging the variable duty cycle waveform of the output of the PDF.

edge arrives before the divider. Likewise, the *DN* (DOWN) signal will produce a one if the divided edge arrives before the reference edge. A simple way to build this circuit, shown in Fig. 17.10, is to use two resettable edge triggered D flip-flops. Both flip-flop inputs are connected to a “one" and will load the value when at the rising edge of their respective clock inputs. One is clocked using the reference while the other is clocked using the divided VCO. After a short delay, the flip flop transfers its input to the output Q, generating an UP or *DN* signal. When both UP and *DN* go to 1, the AND gate resets both flip flops, which means that both UP and *DN* are reset until the arrival of the next edge. In this manner, the charge pump can generate a positive or negative current, for a duration proportional to the time difference between the clock edges. If the edges are close together, a very small output duty cycle current waveform is generated. If the edges are wide apart, a larger duty cycle current waveform is generated.

Unlike the XOR/multiplier, the circuit is sensitive to not only the phase difference, but also the sign of the phase difference. If VCO clock is faster than the reference (higher frequency), then its edges will always arrive earlier, which will activate the *DN* signal which will slow down the VCO. This functionality allows it to function as a frequency detector. The transfer characteristic is derived by observing the average output signal, see Fig. 17.11. Compare Fig. 17.11 with Fig. 17.4. The XOR circuit locks at a phase of $\pi/2$ (quadrature) and only response to a positive phase difference. What happens if the phase difference is negative ? This is certainly a likely scenario, which would result in a flip in the XOR output PD performance. This should scare you, because the slope is now negative, meaning the system becomes a positive feedback system ! Unlike an op-amp that saturates at the rails, the transfer function is periodic, so even if positive feedback pushes the system to the “rails", it will then enter a region when the sign of the transer function is again positive and the system will behave. This behavior is called “cycle slipping" and is only captured by a full non-linear model of the PD.

In the PFD circuit, the transfer function is works for both positive and negative inputs since it has a positive slope for both a positive or negative phase error, meaning that the system will behave linearly even if the phase difference flips polarity. When there’s a frequency difference, the phase of the higher frequency signal is always ramping faster than the lower frequency signal, which is detected by the PDF by generating UP signals, or in other words the PDF is telling the VCO to speed up ! Likewise, if the input signal is faster than the reference, the phase error is always negative, which tells the VCO to slow down !

17.5.4 PFD + Charge Pump

By using the up and down signals to control the charge pump, we can dump or remove charge from the integrating capacitor, and control the VCO, as shown in Fig. 17.12. The functionality the PD and the first integrator are built-in to this block. If neither the up/down signal is activated,

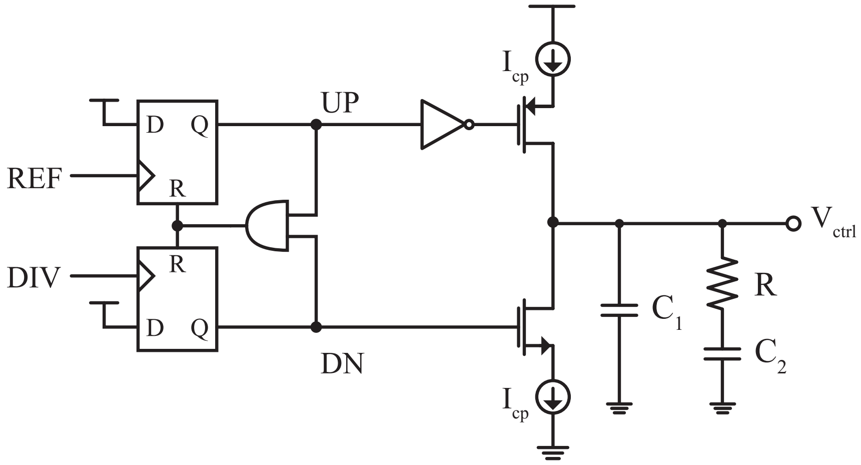


Figure 17.12: Combining the PFD with a charge pump controlled by *UP* and *DN* signals results in the desired loop filter containing a second integrator and a zero for tuning the phase margin.

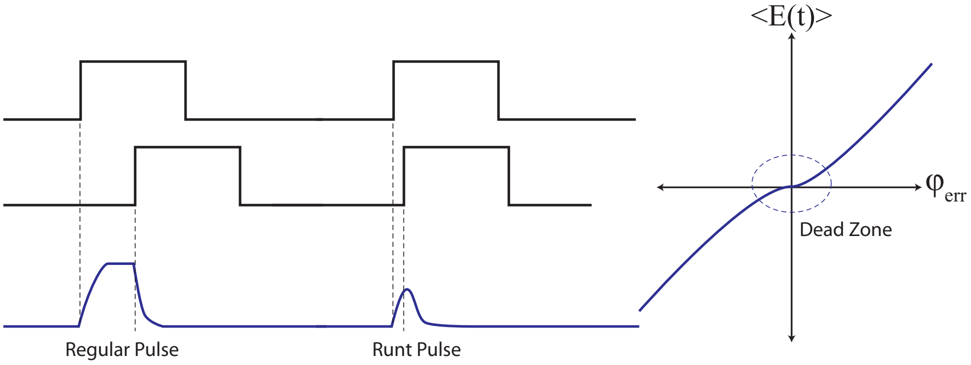


Figure 17.13: The PDF transfer function flattens out for small inputs since the pulse width cannot go below the rise/fall time of the pulses.

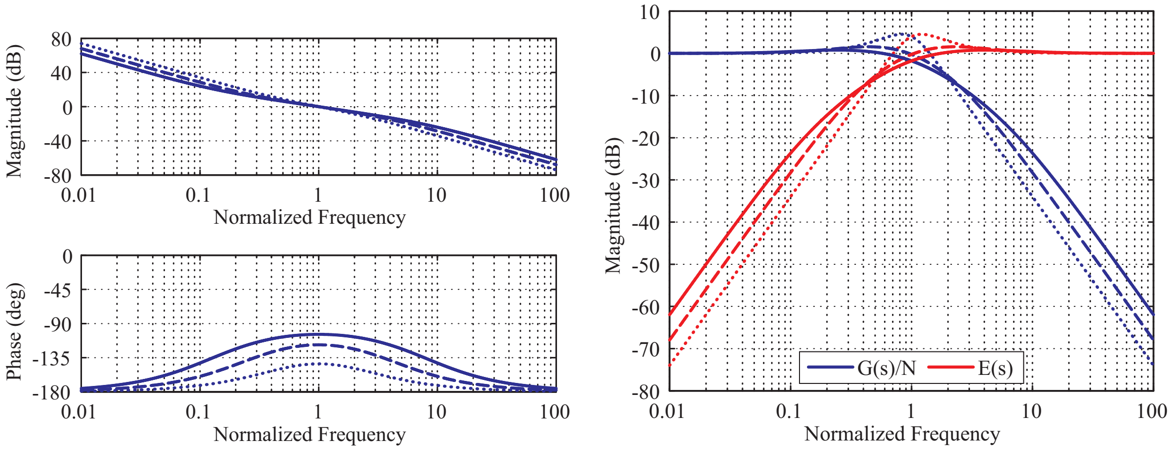
the capacitors hold the charge and the VCO frequency is fixed, which happens in steady state. Any leakage or mismatch between the up/down currents will cause ripples on the control line and therefore reference spurs to be generated. The charge pump devices are sized to minimize the mismatch.

Charge Pump Runt Pulses

Assume that the frequency/phase of the divider and reference are nearly matched so that the phase error is small, as shown in Fig. 17.13. Ideally a shorter and shorter duty cycle signal would be generated, but as the duty cycle approaches the rise time of the pulses, the pulse amplitude will begin to decay, thus lowering the gain of the PDF. We see that the gain of the PDF flattens for small inputs. The solution to this problem is to force the up/down pulses to have a minimum on-time. To produce a small output, therefore, both up and down signals will remain on simultaneously. This can be implemented by using a delay circuit (chain of inverters) after the AND gate.

17.5.5 Loop Gain / Closed Loop Gain

Now let's consider the overall PLL with a charge pump in the loop. Since the gain of the charge pump is $I_{CP} \cdot Z(s)$, and $Z(s)$ is used to realize the loop gain filter, $Z(s) = H(s)$, we now have another



(a) (b)

Figure 17.14: The (a) open- and (b) closed-loop transfer functions of a type II PLL. Two poles at the origin can be observed by the 40 dB/dec roll-off of the transfer function. The closed-loop transfer function exhibits peaking due to the presence of the extra poles/zeros.

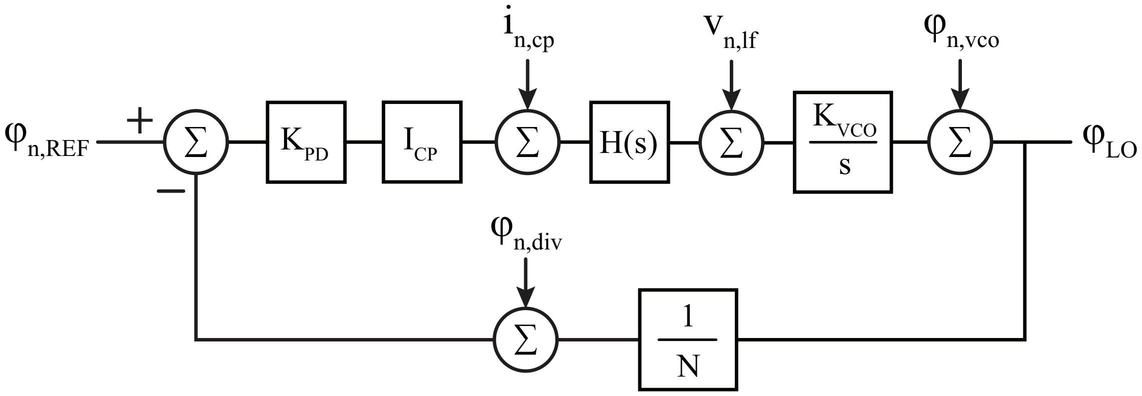


Figure 17.15: PLL linear phase model including source of noise.

knob to tune the loop gain. Earlier we derived the transfer function from the input and VCO output ports. The expression is easily modified to include the charge pump

$$STF = \frac{\Phi_{out}}{\Phi_{ref}} = \frac{\frac{K_{PD}I_{CP}K_{VCO}H(s)}{s}}{1+\frac{K_{PD}I_{CP}K_{VCO}H(s)}{Ns}} = \frac{K_{PD}I_{CP}K_{VCO}H(s)}{s+K_{PD}I_{CP}\frac{K_{VCO}}{N}H(s)} \tag{17.41}$$

As shown in Fig. 17.14, the overall transfer function is now type-II (two integrators) and third-order. There is always some peaking in the transfer curves. Later in Section 17.7 we will approximate the transfer function to see the role of the zero in other overall loop.

17.6 Noise Analysis

In Fig. 17.15 we redraw the linear phase model of the PLL and include various noise sources arising from the VCO (phase noise), the charge pump (current noise), and other sources. For example, divider is often realized by modulating its division ratio between N and $N+1$, which generates an average fractional value and also quantization noise, which we can model as white

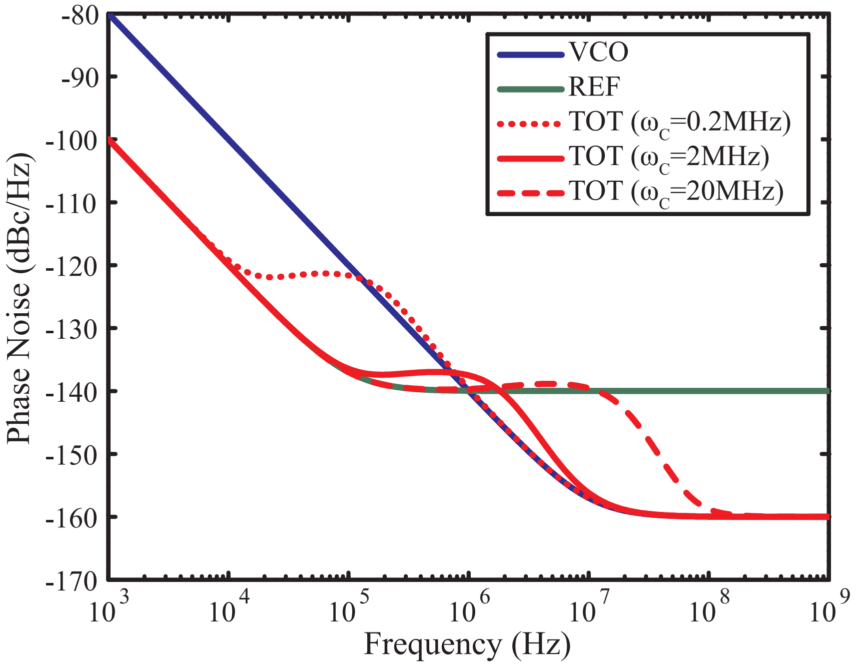


Figure 17.16: The overall phase noise of a PLL broken up into contributions from the VCO and the reference, plotted for three values of the closed-loop bandwidth.

noise for simplicity. Similarly, if the LPF is realized with active circuitry, we can include the noise contribution from active and passive elements using $V_{n,lf}$.

17.6.1 Phase Noise Summary

As before, the transfer function for the VCO noise is given by (HPF)

$$NTF_1 = \frac{\Phi_{out}}{\Phi_{N,VCO}} = \frac{1}{1 + \frac{K_{PD} I_{CP} K_{VCO} H(s)}{Ns}} = \frac{s}{s + \frac{K_{PD} I_{CP} K_{VCO} H(s)}{N}} \qquad (17.42)$$

Since the charge pump is an active circuit, it generates noise and we compute the noise transfer from the CP to the output (LPF)

$$NTF_2 = \frac{\Phi_{out}}{\Phi_{N,CP}} = \frac{H(s)\frac{K_{VCO}}{s}}{1 + \frac{K_{PD} I_{CP} K_{VCO} H(s)}{Ns}} = \frac{K_{VCO} H(s)}{s + \frac{K_{PD} I_{CP} K_{VCO} H(s)}{N}} \qquad (17.43)$$

The phase noise of the VCO is either estimated or simulated and has the well known spectrum that decays like $1/f^3$ due to flicker noise and $1/f^2$ due to the VCO noise shaping (see Sect. **??**). The charge pump noise is highly dependent on its state. If we assume the charge pump is in steady-state, we can use periodic time-varying simulation techniques or calculations to estimate its noise output.

Assuming all the noise sources are independent, the total output noise is given by

$$N_{out}(s) = N_{ref}(s)|STF(s)|^2 + N_{VCO}(s)|NTF_1(s)|^2 + N_{CP}|NTF_2(s)|^2 \qquad (17.44)$$

17.6.2 Phase Noise Spectrum

In Fig. 17.16 we plot the various contributions to the output phase noise. Since the system tracks the reference, it should be obvious that the reference noise appears at the output unattenuated in the pass-band of the PLL. On the other hand, the VCO noise is rejected inside the band. The charge pump should be designed to minimize the in-band noise since its transfer function is low-pass. Recall that any mismatch between the up and down transistors also creates a reference spur, which means the devices should be sized and biased carefully. The total phase noise profile can be characterized into three regions: (1) Reference noise dominates in the PLL bandwidth, (2) the transition band, and (3) outside the loop bandwidth, the “free-running" VCO dominates. We say “free-running" because outside of the loop bandwidth the PLL is not really effective at controlling the VCO, so the VCO noise will appear at the output of the system with a magnitude very much the same as the VCO in free-running mode.

17.7 Loop Dynamics

The transfer function from the input (reference) to the output is given by

$$G(s) = \frac{\phi_{LO}}{\phi_{n,ref}} = \frac{K_{PD} I_{cp} H(s) \frac{K_{VCO}}{s}}{1 + K_{PD} I_{cp} H(s) \frac{K_{VCO}}{s} \frac{1}{N}} \tag{17.45}$$

where $H(s) \approx \frac{1}{sC_2}(1 + s/\omega_z)$. The transfer function is therefore

$$G(s) = \frac{K_{PD} I_{cp} \frac{1}{sC_2}(1 + s/\omega_z) \frac{K_{VCO}}{s}}{1 + K_{PD} I_{cp} \frac{1}{sC_2}(1 + s/\omega_z) \frac{K_{VCO}}{s} \frac{1}{N}} \tag{17.46}$$

$$= \frac{K_{PD} \frac{I_{cp}}{C_2}(1 + s/\omega_z) K_{VCO}}{s^2 + K_{PD} \frac{I_{cp}}{C_2} K_{VCO} \frac{1}{N}(1 + s/\omega_z)} \tag{17.47}$$

Focusing on the denominator, we can put it into standard second-order form

$$D(s) = s^2 + (s/\omega_z) K_{PD} \frac{I_{cp}}{C_2} K_{VCO} \frac{1}{N} + K_{PD} \frac{I_{cp}}{C_2} K_{VCO} \frac{1}{N} \tag{17.48}$$

$$= s^2 + \frac{s\omega_0}{Q} + \omega_0^2 \tag{17.49}$$

17.7.1 Loop Natural Frequency and Quality Factor

By equating the above two equations, we have

$$\omega_0 = \sqrt{K_{PD} \frac{I_{cp}}{C_2} K_{VCO} \frac{1}{N}} \tag{17.50}$$

The amount of ringing in the loop depends on the Q value. The Q is given by

$$\frac{\omega_0}{Q} = \frac{\omega_0^2}{\omega_z} \tag{17.51}$$

or

$$Q = \frac{\omega_z}{\omega_0} \tag{17.52}$$

The location of the zero controls the stability of the loop

17.7.2 Location of Poles

The poles are found easily since it's a second order transfer function

$$s_{1,2} = \frac{\frac{-\omega_0}{Q} \pm \sqrt{\left(\frac{\omega_0}{Q}\right)^2 - 4\omega_0^2}}{2} \tag{17.53}$$

To realize an undamped system (real poles), the $Q < 1/2$. Otherwise there will be peaking in the transfer function, also known as jitter peaking (jitter is the time domain equivalent of phase noise).

"Critically Damped" System

If we take $Q = 1/2$, we have two poles at

$$s_{1,2} = \frac{-\omega_0}{2Q} = -\omega_0 \tag{17.54}$$

The location of the zero is $\omega_z = Q\omega_0 = 0.5\omega_0$. The overall transfer function is given by

$$G(s) = G(0)\frac{1+2s/\omega_0}{(1+s/\omega_0)^2} \tag{17.55}$$

Due to the zero, the system still overshoots slightly and the loop bandwidth is given by $\omega_{-3dB} \approx 2.5\omega_0$.

Underdamped System

To ensure a damped response, one may choose $Q \ll 1$, say $Q = 0.1$, which implies a very low frequency zero, $\omega_z = Q\omega_0$. This requires a large capacitor! A useful approximation for a low Q system is the following

$$s^2 + \frac{s\omega_0}{Q} + \omega_0^2 \approx (s + \frac{\omega_0}{Q})(s + Q\omega_0) \tag{17.56}$$

The utility of this approximation is that the second pole is at the zero frequency $\omega_z = Q\omega_0$, which cancels the zero in the transfer function

$$G(s) = \frac{K_{VCO}K_{PD}\frac{I_{cp}}{C_2}(1+s/\omega_z)}{(s+\frac{\omega_0}{Q})(s+\omega_z)} \tag{17.57}$$

$$G(s) = \frac{K_{VCO}K_{PD}\frac{I_{cp}}{\omega_z C_2}(1+s/\omega_z)}{(s+\frac{\omega_0}{Q})(1+s/\omega_z)} \tag{17.58}$$

The overall transfer function simplifies and acts like a single pole system

$$G(s) = \frac{K_{VCO}K_{PD}\frac{I_{cp}}{\omega_z C_2}}{(s+\frac{\omega_0}{Q})} \tag{17.59}$$

The pole is given by

$$\omega_{-3dB} = \frac{\omega_0}{Q} = \frac{1}{Q}\sqrt{K_{PD}\frac{I_{cp}}{C_2}K_{VCO}\frac{1}{N}} \tag{17.60}$$

Keep in mind that this is an approximation and in reality the pole/zero don't cancel exactly. Also we neglected the higher order pole of the lead/lag filter. There are more poles in the system as well, so simulate everything to be sure !

17.7.3 Choice of Loop Bandwidth

Even though we modeled the PLL as a continuous time system, in reality most implementations are a discrete time system. This is due to the fact that the PDF compares edges of the reference to the divided clock, and only generates an error signal once per reference edge. One implication of the discrete time system is that the bandwidth must be smaller than the reference frequency, typically 1/10'th of the reference is a popular choice. XTAL oscillators are available up to a few hundred megahertz. In an Integer-N architecture, the reference frequency is set to the channel spacing, which means the loop dynamics are controlled by the reference frequency. Accurate synthesizers are therefore slow. A fractional-N synthesizer in theory decouples the choice of reference frequency from the resolution of the PLL, but in practice fractional spurs force low bandwidths. Realization of a high bandwidth fractional-N PLL is an active research topic.

17.8 Simulation Tools

These equations are a good starting point in the design of a PLL. The next step is a system level simulation in matlab or an analysis with the complete transfer function. Next, you should remind yourself that our linear model is an approximation. A simulation framework that can model the actual dynamics, including the non-linearity, is very important. Full SPICE level simulation is too slow (hours to days) for design, but is a must for verification. Verilog-A is a great choice to model the PLL, and even include some of the blocks at the transistor level.

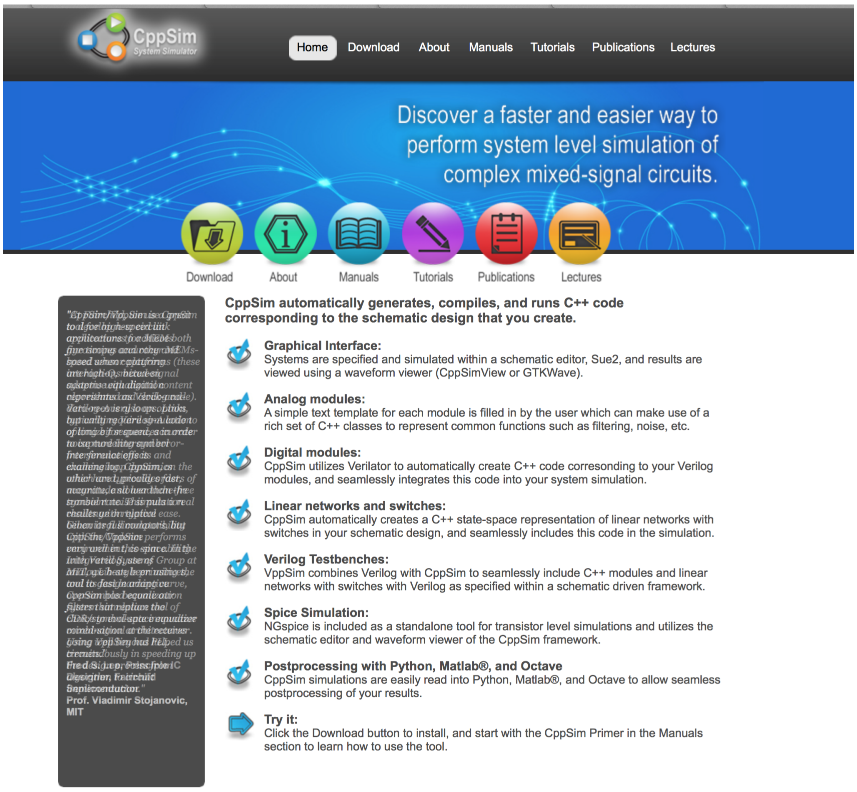
17.8.1 CppSim

A great free tool that runs very fast (C++ based platform) and includes support for verilog components is available and highly recommended: `cppsim.org`, see Fig. 17.17.

17.9 Charge Pump Realizations

1. Simple current mirror
2. Op-amp to match and positive feedback loop

17.10 Conclusion



"[illegible]"
Prof. Vladimir Stojanovic, MIT

CppSim automatically generates, compiles, and runs C++ code corresponding to the schematic design that you create.

Graphical Interface:
Systems are specified and simulated within a schematic editor, Sue2, and results are viewed using a waveform viewer (CppSimView or GTKWave).

Analog modules:
A simple text template for each module is filled in by the user which can make use of a rich set of C++ classes to represent common functions such as filtering, noise, etc.

Digital modules:
CppSim utilizes Verilator to automatically create C++ code corresonding to your Verilog modules, and seamlessly integrates this code into your system simulation.

Linear networks and switches:
CppSim automatically creates a C++ state-space representation of linear networks with switches in your schematic design, and seamlessly includes this code in the simulation.

Verilog Testbenches:
VppSim combines Verilog with CppSim to seamlessly include C++ modules and linear networks with switches with Verilog as specified within a schematic driven framework.

Spice Simulation:
NGspice is included as a standalone tool for transistor level simulations and utilizes the schematic editor and waveform viewer of the CppSim framework.

Postprocessing with Python, Matlab®, and Octave
CppSim simulations are easily read into Python, Matlab®, and Octave to allow seamless postprocessing of your results.

Try it:
Click the Download button to install, and start with the CppSim Primer in the Manuals section to learn how to use the tool.

Figure 17.17: CPPSim is a graphical simulation framework optimized for mixed-mode systems that include analog / RF blocks as behavior models and digital systems using verilog or custom C++ modules. A full PLL can be simulated much faster than a full transistor level SPICE circuit.