John W. M. Rogers • Calvin Plett

RADIO FREQUENCY INTEGRATED CIRCUIT DESIGN

SECOND EDITION

Radio Frequency Integrated Circuit Design

Second Edition

For a list of recent related titles in the *Artech House Microwave Library*,
please turn to the back of this book.

Radio Frequency Integrated Circuit Design

Second Edition

John W. M. Rogers
Calvin Plett

ARTECH
HOUSE
BOSTON | LONDON
artechhouse.com

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the U.S. Library of Congress.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN-13: 978-1-60783-979-8

Cover design by Igor Valdman

685 Canton Street
Norwood, MA 02062

10 9 8 7 6 5 4 3 2 1

Contents

Foreword to the First Edition

I enjoyed reading this book for a number of reasons. One reason is that it addresses high-speed analog design in the context of microwave issues. This is an advanced level book, which should follow courses in basic circuits and transmission lines. Most analog integrated circuit designers in the past worked on applications at a low enough frequency that microwave issues did not arise. As a consequence, they were adept at lumped parameter circuits and often not comfortable with circuits where waves travel in space. However, in order to design radio frequency (RF) communications integrated circuits (IC) in the gigahertz range, one must deal with transmission lines at chip interfaces and where interconnections on chips are far apart. Also, impedance matching is addressed, which is a topic that arises most often in microwave circuits. In my career, there has been a gap in comprehension between analog low-frequency designers and microwave designers. Often, similar issues were dealt with in two different languages. Although this book is more firmly based in lumped-element analog circuit design, it is nice to see that microwave knowledge is brought in where necessary.

Too many analog circuit books in the past have concentrated first on the circuit side rather than on basic theory behind their application in communications. The circuits usually used have evolved through experience, without a satisfying intellectual theme in describing them. Why a given circuit works best can be subtle; often these circuits are chosen only through experience. For this reason, I am happy that the book begins first with topics that require an intellectual approach—noise, linearity and filtering, and technology issues. I am particularly happy with how linearity is introduced (power series). In the rest of the book it is then shown, with specific circuits and numerical examples, how linearity and noise arise.

In the latter part of the book, the RF circuits analyzed are ones that experience has shown to be good. Concentration is on bipolar circuits, not metal oxide semiconductors (MOS). Bipolar still has many advantages at high frequencies. The depth with which design issues are addressed would not be possible if similar MOS coverage was attempted. However, there might be room for a similar book, which concentrates on MOS.

In this book, there is much detailed academic exploration of some important high-frequency RF bipolar ICs. One might ask if this is important in design for application, and the answer is yes. To understand why, one must appreciate the central role of analog circuit simulators in the design of such circuits. At the beginning of my career (around 1955–1960) discrete circuits were large enough that good circuit topologies could be picked out by breadboarding with the actual parts themselves.

This worked fairly well with some analog circuits at audio frequencies, but failed completely in the progression to integrated circuits.

In high-speed IC design nowadays, the computer-based circuit simulator is crucial. Such simulation is important at four levels. The first level is the use of simplified models of the circuit elements (idealized transistors, capacitors, and inductors). The use of such models allows one to pick out good topologies and eliminate bad ones. This is not done well with just paper analysis because it will miss key factors, such as the complexities of the transistor, particularly nonlinearities and bias and signal interaction effects. Exploration of topologies with the aid of a circuit simulator is necessary. The simulator is useful for quick iteration of proposed circuits with simplified models to show any fundamental problems with a proposed circuit. This brings out the influence of model problems on circuit performance. This first level of simulation may be avoided if the best topology, known through experience, is picked at the start.

The second level of simulation is where the models used are representative of the type of fabrication technology being used. However we do not yet use specific numbers from the specific fabrication process, and make an educated approximation to likely parasitic capacitances. Simulation at this level can be used to hone in on good values for circuit parameters for a given topology, before the final fabrication process is available. Before the simulation begins, detailed preliminary analysis at the level of this book is possible and many parameters can be wisely chosen before simulation begins, greatly shortening the design process and the required number of iterations. Thus the analysis should focus on topics that arise, given a typical fabrication process. I believe that this has been done well here, and the authors through scholarly work and real design experience, have chosen key circuits and topics.

The third level of design is where a link with a proprietary industrial process has been made, and good simulator models are supplied for the process. The circuit is laid out in the proprietary process and a simulation is done, including estimates of parasitic capacitances from interconnections and detailed models of the elements used.

The incorporation of the proprietary models in the simulation of the circuit is necessary because when the IC is laid out in the actual process, fabrication of the result must be successful to the highest possible degree. This is because fabrication and testing is extremely expensive. Any failure can result in the necessity to change the design, which requires further fabrication and retesting, and causes a delay in getting the product to the market.

The fourth design level is the comparison of the circuit behavior predicted from simulation with that of measurements of the actual circuit. Discrepancies must be explained. These may be from design errors or from inadequacies in the models, which are uncovered by the experimental result. These model inadequacies, when corrected, may result in further simulation, which causes the circuit design and layout to be refined with further fabrication.

This discussion has served to bring attention to the central role that computer simulation has in the design of integrated RF circuits, and the accompanying importance of circuit analysis as presented in this book. Such detailed analysis may save

money by facilitating the early success of applications. This book can be beneficial to designers—or to those less focused on specific design—to recognize key constraints in the area, with faith (justified, I believe) that the book is a correct picture of the reality of high speed RF communications circuit design.

Miles A. Copeland
IEEE Fellow
Professor Emeritus
Carleton University
Department of Electronics
Ottawa, Ontario, Canada
April 2003

Preface

For this second edition of this book, we have added significantly more information about radio frequency integrated circuit (RFIC) design using complementary metal oxide semiconductor (CMOS) transistors, whereas the first edition strongly emphasized design with bipolar transistors. Since 2003, when the first edition of this book was published, CMOS has grown in popularity both in industry and in academia, and every year in our own RFIC design courses we have been putting more emphasis on CMOS. Although many RFIC design principles are the same whether using bipolar transistors or CMOS transistors, there are some important differences. In this edition of the book, we have added the relevant equations for CMOS circuits and we have added complete CMOS-based design examples for each of the major radio frequency building blocks.

In addition, we have added a chapter on system issues (Chapter 3) and a chapter on frequency synthesizer design (Chapter 10). Both of these topics could have a complete book written about them, so our coverage is intended to include what we consider to be the most important points.

Another major reason for this new edition is to correct some of the numerous mistakes that unfortunately found their way into the first edition. However, since there will likely still be mistakes, we will provide a list of errors on a Web page (http://www.doe.carleton.ca/cp/rfic2_errata.html) and we encourage you to make use of it and to inform us of any additional errors not yet listed on this page.

Acknowledgments

The first edition of this book evolved out of a number of documents including technical papers, course notes, and various theses. We decided that we would organize some of the research we and many others had been doing, and turn it into a manuscript that would serve as a comprehensive text for engineers interested in learning about radio frequency integrated circuits (RFIC). We have focused mainly on bipolar technology in the text, but since many techniques in RFICs are independent of technology, we hope that designers working with other technologies will also find much of the text useful. We have tried very hard to identify and exterminate bugs and errors from the text. Undoubtedly there are still many remaining, so we ask you, the reader, for your understanding. Please feel free to contact us with your comments. We hope that these pages add to your understanding of the subject.

Nobody undertakes a project like this without support on a number of levels, and there are many people that we need to thank. Professors Miles Copeland and Garry Tarr provided technical guidance and editing. We would like to thank David Moore for his input and consultation on many aspects of RFIC design. David, we have tried to add some of your wisdom to these pages. Thanks also go to Dave Rahn and Steve Kovacic, who have both contributed to our research efforts in a variety of ways. We would like to thank Sandi Plett who tirelessly edited chapters, provided formatting, and helped beat the word processor into submission. She did more than anybody except the authors to make this project happen. We would also like to thank a number of graduate students, alumni, and colleagues who have helped us with our understanding of RFICs over the years. This list includes but is not limited to: Neric Fong, Bill Toole, José Macedo, Sundus Kubba, Leonard Dauphinee, Rony Amaya, John J. Nisbet, Sorin Voinegescu, John Long, Tom Smy, Walt Bax, Brian Robar, Richard Griffith, Huges Lafontaine, Ash Swaminathan, Jugnu Ojha, George Khoury, Mark Cloutier, John Peirce, Bill Bereza, and Martin Snelgrove.

For the second edition of the book, we would like to acknowledge the help from research collaborators, colleagues, and students. This list includes, but is not limited to: Jim Wight, Foster Dai, Mark Cavin, Oliver Werther, Angelika Schneider, Robert Renninger, John Marcincavage, Robson Nunes de Lima, Gina Zhou, Peter Popplewell, Victor Karam, Kobe Situ, Jorge Aguirre, John Danson, Justin Abbott, James Chiu, Travis Lovitt, Kimia Ansari, Omid Salehi-Abari, Saber Amini, Samira Bashiri, Jerry Lam, Ché Knisely, Andrea Liao, Yasser Soliman, Steve Knox, and Steve Penney.

CHAPTER 1

Introduction to Communications Circuits

1.1 Introduction

Radio frequency integrated circuit (RFIC) design is an exciting area for research or product development. Technologies are constantly being improved, and as they are, circuits formerly implemented as discrete solutions can now be integrated onto a single chip. In addition to widely used applications such as cordless phones and cell phones, new applications continue to emerge. Examples of new products requiring RFICs are wireless local area networks (WLAN), keyless entry for cars, wireless toll collection, global positioning system (GPS) navigation, radio frequency identification (RFID) tags, asset tracking, remote sensing, ultra wideband (UWB) radios for high data rate personal area networks (PANs), mobile television reception using standards such as digital video broadcast handheld (DVB-H) or digital video broadcast terrestrial (DVB-T), and tuners in cable modems. Thus, the market is expanding, and with each new application there are unique challenges for the designers to overcome. As a result, the field of RFIC design should have an abundance of products to keep designers entertained for years to come.

This huge increase in interest in radio frequency (RF) communications has resulted in an effort to provide components and complete systems on an integrated circuit (IC). There has been much research aimed at putting a complete radio on one chip. Since complementary metal oxide semiconductor (CMOS) is required for the digital signal processing (DSP) in the back end, much of this effort has been devoted to radios in CMOS technologies [1–3]. However, bipolar design continues to be used in the industry because of its higher performance. CMOS traditionally had the advantage of lower production cost, but this is becoming less true as technology dimensions become smaller. In the future, both of these technologies will probably be replaced by radically different technologies. In any case, as long as people want to communicate, engineers will still be building radios, and contrary to popular belief, most of the design concepts in RFIC design are applicable regardless of what technology is used to implement them.

The objective of a radio is to transmit and receive signals between the source and destination with an acceptable quality and without incurring a high cost. From the user's point of view, quality can be perceived as information being passed from source to destination without the addition of noticeable noise or distortion. From a more technical point of view, quality is often measured in terms of bit error rate, and acceptable quality might be to experience less than one error in every million bits. Cost can be seen as the price of the communications equipment or the need to replace or recharge batteries. Low cost implies simple circuits to minimize circuit area and low power dissipation to maximize battery life.

1.2 Lower Frequency Analog Design and Microwave Design Versus Radio-Frequency Integrated Circuit Design

Radio-frequency integrated circuit design has borrowed from both analog design techniques, used at lower frequencies [4, 5], and high frequency design techniques, making use of microwave theory [6, 7]. The most fundamental difference between low frequency analog and microwave design is that in microwave design, transmission line concepts are important, while in low-frequency analog design, they are not. This will have implications on the choice of impedance levels, as well as how signal size, noise, and distortion are described.

On-chip dimensions are small, so even at RF frequencies (0.1–60 GHz), transistors and other devices may not need to be connected by transmission lines (i.e., the lengths of the interconnects may not be a significant fraction of a wavelength). However, at the chip boundaries, or when traversing a significant fraction of a wavelength on-chip, transmission line theory becomes very important. Thus, on chip we can usually make use of analog design concepts, although, in practice, microwave design concepts are often used. Where the chip interfaces with the outside world, we must treat it like a microwave circuit.

1.2.1 Impedance Levels for Microwave and Low-Frequency Analog Design

In low-frequency analog design, input impedance is usually very high (ideally infinity) while output impedance is low (ideally zero). For example, an operational amplifier can be used as a buffer because its high input impedance does not affect the circuit to which it is connected, and its low output impedance can drive a measurement device efficiently. The freedom to choose arbitrary impedance levels provides advantages in that circuits can drive or be driven by an impedance that best suits them. On the other hand, if circuits are connected using transmission lines, then these circuits are usually designed to have an input and output impedance that match the characteristic impedance of the transmission line.

1.2.2 Units for Microwave and Low-Frequency Analog Design

Signal, noise, and distortion levels are also described differently in low frequency analog versus microwave design. In microwave circuits, power is usually used to describe signals, noise, or distortion with the typical unit of measure being decibels above 1 milliwatt (dBm). However, in analog circuits, since infinite or zero impedance is allowed, power levels are meaningless, so voltages and currents are usually chosen to describe the signal levels. Voltage and current levels are expressed as peak, peak-to-peak, or root-mean-square (rms). Power in dBm, P_{dBm}, can be related to the power in watts, P_{watt}, as shown in (1.1) and Table 1.1, where voltages are assumed to be across 50 .

$$P_{\text{dbm}} = 10\log_{10}\left(\frac{P_{\text{watt}}}{1\,\text{mW}}\right) \tag{1.1}$$

Table 1.1 Power Relationships

v_{pp}	v_{rms}	P_{watt} (50Ω)	P_{dBm} (50Ω)
1 nV	0.3536 nV	2.5 10 ^{21}	176
1 μV	0.3536 μV	2.5 10 ^{15}	116
1 mV	353.6 mV	2.5 nW	56
10 mV	3.536 mV	250 nW	36
100 mV	35.36 mV	25 μW	16
632.4 mV	223.6 mV	1 mW	0
1V	353.6 mV	2.5 mW	+4
10V	3.536V	250 mW	+24

Assuming a sinusoidal voltage waveform, P_{watt} is given by

$$P_{\text{watt}} = \frac{v_{\text{rms}}^2}{R} \tag{1.2}$$

where R is the resistance the voltage is developed across. Note also that v_{rms} can be related to the peak-to-peak voltage v_{pp} by

$$v_{\text{rms}} = \frac{v_{\text{pp}}}{2\sqrt{2}} \tag{1.3}$$

Similarly, noise in analog signals is often defined in terms of volts or amperes, while in microwave it will be in terms of dBm. Noise is usually represented as noise density per hertz of bandwidth. In analog circuits, noise is specified as squared volts per hertz, or volts per square root of hertz. In microwave circuits, the usual measure of noise is dBm/Hz or noise figure, which is defined as the reduction in signal-to-noise ratio caused by the addition of the noise.

In both analog and microwave circuits, an effect of nonlinearity is the appearance of harmonic distortion components or intermodulation distortion components, often at new frequencies. In low-frequency analog circuits, this is often described by the ratio of the distortion components compared to the fundamental components. In microwave circuits, the tendency is to describe distortion by gain compression (power level where the gain is reduced due to nonlinearity) or the third-order intercept point (IP3).

Noise and linearity are discussed in detail in Chapters 2 and 3. A summary of low frequency analog and microwave design is shown in Table 1.2.

Table 1.2 Comparison of Analog and Microwave Design

Parameter	*Analog Design (Most Often Used On-Chip)*	*Microwave Design (Most Often Used at Chip Boundaries and Pins)*
Impedance	Z_{in} ∞ Z_{out} 0	Z_{in} 50 Z_{out} 50
Signals	Voltage, current, often peak or peak-to-peak	Power, often dBm
Noise	$\text{nV}/\sqrt{\text{Hz}}$	Noise factor *F*, noise figure *NF*
Nonlinearity	Harmonic distortion, intermodulation, clipping	Third-order intercept point IP3, 1-dB compression

In summary, as stated earlier, RF design is a combination of techniques used in low-frequency analog design, and techniques used in traditional microwave design. On chip if distances between components are small, some analog design concepts are used. That is, transmission lines are generally not used and convenient impedance levels are chosen. For example, impedances higher than 50 may be used to save power, or low impedances may be chosen to achieve higher bandwidth in the presence of capacitance. It should be noted, however, that the low frequency technique of aiming for infinite input impedance and zero output impedance is not possible at radio frequencies. Design is often done using small signal models, and internal signals are often described using voltage or current as is common for low frequency analog design. Conversely, at chip boundaries and for components separated by long distances, transmission lines are used and impedances are chosen to optimize power transfer or minimize noise according to standard microwave techniques.

1.3 Radio-Frequency Integrated Circuits Used in a Communications Transceiver

A typical block diagram of most of the major circuit blocks that make up a typical superheterodyne communications transceiver is shown in Figure 1.1. Many aspects of this transceiver are common to all transceivers.

This transceiver has a transmit side (Tx) and a receive side (Rx), which are connected to the antenna through a duplexer that can be realized as a switch or a filter, depending on the communications standard being followed. The input preselection filter takes the broad spectrum of signals coming from the antenna and removes the signals that are not in the band of interest. This may be required to prevent overloading of the low-noise amplifier (LNA) by out-of-band signals. The

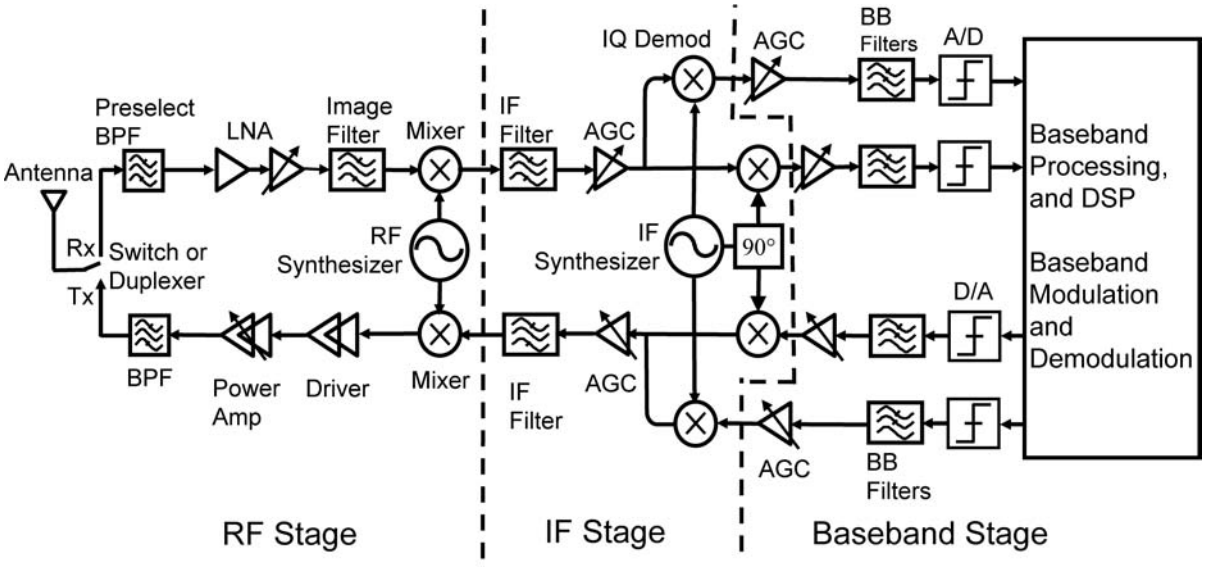

Figure 1.1 Typical transceiver block diagram.

LNA amplifies the input signal without adding much noise. The input signal can be very weak, so the first thing to do is strengthen the signal without corrupting it. As a result, noise added in later stages will be of less importance. The image filter that follows the LNA removes out-of-band signals and noise before the signal enters the mixer. The mixer translates the input radio frequency signal down to the intermediate frequency, since filtering, as well as circuit design, become much easier at lower frequencies for a multitude of reasons. The other input to the mixer is the local oscillator (LO) signal provided by a voltage-controlled oscillator inside a frequency synthesizer. The desired output of the mixer will be the difference between the LO frequency and the RF frequency.

At the input of the radio, there may be many different channels or frequency bands. The LO frequency is adjusted so that the desired RF channel or frequency band is mixed down to the same intermediate frequency (IF) frequency in all cases. The IF stage then provides channel filtering at this frequency to remove the unwanted channels. The IF stage provides further amplification and automatic gain control (AGC) to bring the signal to a specific amplitude level before the signal is passed on to the back end of the receiver. It will ultimately be converted into bits (most modern communications systems use digital modulation schemes) that could represent voice, video, data, and so forth, through the use of an A/D converter.

On the transmit side, the back-end digital signal is used to modulate the carrier in the IF stage. A mixer converts the modulated signal and IF carrier up to the desired RF frequency. A frequency synthesizer provides the other mixer input. Since the RF carrier and associated modulated data may have to be transmitted over large distances through lossy media (e.g., air, cable, or fiber), a power amplifier (PA) must be used to increase the signal power. Typically, the power level is increased from the milliwatt range to a level in the range of hundreds of milliwatts, to watts, depending on the particular application. A low pass filter after the PA removes any harmonics produced by the PA to prevent them from also being transmitted. Radio architectures will be studied in much more detail in Chapter 3.

1.4 Overview

We will spend the rest of this book trying to convey the various design constraints of the RF building blocks mentioned in the previous sections. Components are designed with consideration to the main problems of frequency response, gain, stability, noise, distortion (nonlinearity), impedance matching, and power dissipation. Dealing with design constraints is what keeps the RFIC designer employed.

The focus of this book will be how to design and build the major circuit blocks that make up the RF portion of a radio using an IC technology. To that end, block level performance specifications are described in Chapter 2 and architecture considerations are discussed in Chapter 3. A brief overview of IC technologies and transistor performance is given in Chapter 4. Various methods of impedance matching, which is very important at chip boundaries and for some interconnections of circuits on chip, will be discussed in Chapter 5. The realization and limitations of passive circuit components in an IC technology will be discussed in Chapter 6. Chapters 7, 8, 9, and 11 will be devoted to individual circuit blocks such as LNAs

mixers, voltage-controlled oscillators (VCOs), and power amplifiers. Synthesizers will be discussed in Chapter 10.

References

[1] Lee, T. H., *The Design of CMOS Radio Frequency Integrated Circuits*, 2nd ed., Cambridge, U.K.: Cambridge University Press, 2004.

[2] Razavi, B., *RF Microelectronics*, Upper Saddle River, NJ: Prentice-Hall, 1998.

[3] Crols, J., and M. Steyaert, *CMOS Wireless Transceiver Design*, Dordrecht, the Netherlands: Kluwer Academic Publishers, 1997.

[4] Gray, P. R., et al., *Analysis and Design of Analog Integrated Circuits*, 4th ed., John Wiley & Sons, 2001.

[5] Johns, D. A., and K. Martin, *Analog Integrated Circuit Design*, New York: John Wiley & Sons, 1997.

[6] Gonzalez, G., *Microwave Transistor Amplifiers Analysis and Design*, 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 1997.

[7] Pozar, D. M., *Microwave Engineering*, 2nd ed., New York: John Wiley & Sons, 1998.

CHAPTER 2

Issues in RFIC Design: Noise, Linearity, and Signals

2.1 Introduction

In this chapter we will have a brief look at some general issues in RF circuit design. The nonidealities we will consider include noise and nonlinearity. An ideal circuit, for example an amplifier, produces a perfect copy of the input signal at the output. In a real circuit, the amplifier will add both noise and distortion to that waveform. Noise, which is present in all resistors and active devices, limits the minimum detectable signal in a radio. At the other amplitude extreme, nonlinearities in the circuit blocks will cause the output signal to become distorted, limiting the maximum signal amplitude.

At the system level, specifications for linearity and noise, as well as many other parameters, must be determined before the circuit can be designed. In this chapter, before we look at circuit design, we will look at some of these system issues in more detail. In order to design radio frequency integrated circuits with realistic specifications, we need to understand the impact of noise on minimum detectable signals and the effect of nonlinearity on distortion. Knowledge of noise floors and distortion will be used to understand the requirements for circuit parameters. Finally, the chapter will conclude by studying some common digital modulations used in communications. In many cases the modulation used will determine the performance levels required by the various RFIC blocks to make the radio work properly.

2.2 Noise

Signal detection is more difficult in the presence of noise. In addition to the desired signal, the receiver also picks up noise from the rest of the universe. Any matter above zero Kelvin contains thermal energy. This thermal energy moves atoms and electrons around in a random way, leading to random currents in circuits, which are also seen as noise. Noise can also come from manmade sources such as microwave ovens, cell phones, pagers, radio antennas, and so forth. Circuit designers are mostly concerned with how much noise is being added by the circuits in the transceiver. At the input to the receiver, there will be some noise power present, which defines the noise floor. The minimum detectable signal must be higher than the noise floor by some signal-to-noise ratio (SNR) to detect signals reliably and to compensate for additional noise added by circuitry. These concepts will be described in the following sections.

To find the total noise due to a number of sources, the relationship of the sources with each other has to be considered. The most common assumption is that all noise sources are random and have no relationship with each other, so they are said to be uncorrelated. In such a case, noise power is added instead of noise voltage. Similarly, if noise at different frequencies is uncorrelated, noise power is added. We note that signals, like noise, can also be uncorrelated, for example, signals at different unrelated frequencies. In such a case, one finds the total output signal by adding the powers. On the other hand, if two sources are correlated, the voltages can be added. As an example, correlated noise is seen at the outputs of two separate paths that have the same origin.

2.2.1 Thermal Noise

One of the most common noise sources in a circuit is a resistor. Noise in resistors is generated by thermal energy causing random electron motion [1–3]. The thermal noise spectral density in a resistor is given by:

$$N_{\text{resistor}} = 4kTR \tag{2.1}$$

where T is the temperature in Kelvin of the resistor, k is Boltzmann's constant (1.38 10^{-23} joule/K) and R is the value of the resistor. Noise power spectral density has the units of V^2/Hz (power spectral density). To determine how much power a resistor produces in a finite bandwidth, simply multiply (2.1) by the bandwidth of interest f:

$$v_n^2 = 4kTR \quad f \tag{2.2}$$

where v_n is the rms value of the noise voltage in the bandwidth f. This can also be written equivalently as a noise current rather than a noise voltage:

$$i_n^2 = \frac{4kT \quad f}{R} \tag{2.3}$$

Thermal noise is white noise, meaning it has a constant power spectral density with respect to frequency (valid up to approximately 6,000 GHz) [4]. The model for noise in a resistor is shown in Figure 2.1.

2.2.2 Available Noise Power

Maximum power is transferred to the load when R_{LOAD} is equal to R. Then v_o is equal to $v_n/2$. The output power spectral density P_o is then given by:

$$P_o = \frac{v_o^2}{R} = \frac{v_n^2}{4R} = kT \tag{2.4}$$

Thus, available power is kT, independent of resistor size. Note that kT is in watts per hertz, which is a power density. To get total power out P_{out} in watts, multiply by the bandwidth, with the result that:

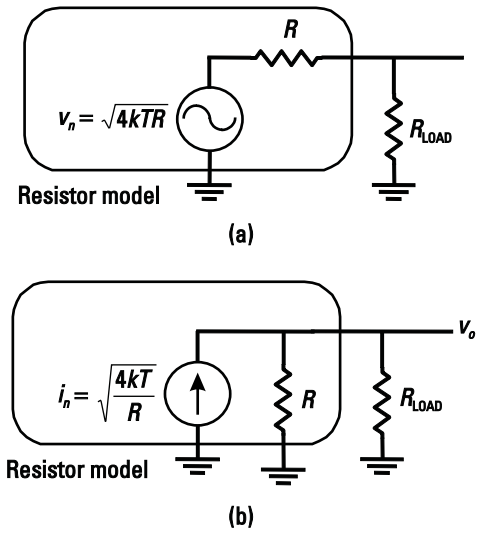

Figure 2.1 Resistor noise model: (a) with a voltage source and (b) with a current source.

$$P_{\text{out}} = kTB \tag{2.5}$$

2.2.3 Available Power from Antenna

The noise from an antenna can be modeled as a resistor [5]. Thus, as in Section 2.2.2, the available power from an antenna is given by:

$$P_{\text{available}} = kT = 4 \quad 10^{-21}\,\text{W/Hz} \tag{2.6}$$

at T = 290K, or in dBm/Hz:

$$P_{\text{available}} = 10\log_{10}\left(\frac{4 \quad 10^{-21}}{1 \quad 10^{-3}}\right) = \quad 174\ \text{dBm/Hz} \tag{2.7}$$

Using 290K as the temperature of the resistor modeling the antenna is appropriate for cell phone applications where the antenna is pointed at the horizon. However, if the antenna were pointed at the sky, the equivalent noise temperature would be much lower—typically 50K [6].

For any receiver required to receive a given bandwidth of signal, the minimum detectable signal can now be determined. As can be seen from (2.5), the noise floor depends on the bandwidth. For example, with a bandwidth of 200 kHz, the noise floor is:

$$\text{Noise Floor} = kTB = 4 \quad 10^{-21} \quad 200{,}000 = 8 \quad 10^{-16}\,\text{W} \tag{2.8}$$

More commonly, the noise floor would be expressed in dBm, as in the following for the same example as before:

$$\text{Noise Floor} = \quad 174\ \text{dBm/Hz} + 10\log_{10}(200{,}000) = \quad 121\ \text{dBm} \tag{2.9}$$

Thus, we can now also formally define signal-to-noise ratio. If the signal has a power of S then the SNR is:

$$\text{SNR} = \frac{S}{\text{Noise Floor}} \tag{2.10}$$

Thus, if the electronics added no noise and if the detector required a signal-to-noise ratio (SNR) of 0 dB, then a signal at 121 dBm could just be detected. The minimum detectable signal in a receiver is also referred to as the receiver sensitivity. However, the SNR required to detect bits reliably, (e.g., BER = 10^{-3}) is typically not 0 dB. The actual required SNR depends on a variety of factors, such as bit rate, energy per bit, IF filter bandwidth, detection method (e.g., synchronous or not), interference levels, and so forth. Such calculations are the topics of a digital communications course [6, 7] and will also be discussed in Section 2.4.1. Typical results for a bit error rate of 10^{-3} is about 7 dB for QPSK, about 12 dB for 16 QAM, and about 17 dB for 64 QAM, though often higher numbers are quoted to leave a safety margin. It should be noted that for data transmission, lower BER is often required (e.g., 10^{-6} resulting in an SNR requirement of 11 dB or more for QPSK). Thus, the input signal level must be above the noise floor level by at least this amount. Consequently, the minimum detectable signal level in a 200-kHz bandwidth is more like 114 to 110 dBm (assuming no noise is added by the electronics).

2.2.4 The Concept of Noise Figure

Noise added by electronics will directly add to the noise from the input. Thus, for reliable detection, the previously calculated minimum detectable signal level must be modified to include the noise from the active circuitry. Noise from the electronics is described by noise factor F, which is a measure of how much the signal-to-noise ratio is degraded through the system. By noting that:

$$S_o = G \times S_i \tag{2.11}$$

where S_i is the input signal power, S_o is the output signal power, and G is the power gain S_o/S_i. We derive the following for noise factor:

$$F = \frac{\text{SNR}_i}{\text{SNR}_o} = \frac{S_i / N_{i(\text{source})}}{S_o / N_{o(\text{total})}} = \frac{S_i / N_i}{(S \times G) / N_{o(\text{total})}} = \frac{N_{o(\text{total})}}{G \times N_{i(\text{source})}} \tag{2.12}$$

where $N_{o(\text{total})}$ is the total noise at the output, $N_{o(\text{source})}$ is the noise at the output originating from the source, and $N_{o(\text{added})}$ is the noise at the output added by the electronic circuitry. Noting that:

$$N_{o(\text{total})} = N_{o(\text{source})} + N_{o(\text{added})} \tag{2.13}$$

noise factor can be written in several useful alternative forms:

$$F = \frac{N_{o(\text{total})}}{G \times N_{i(\text{source})}} = \frac{N_{o(\text{total})}}{N_{o(\text{source})}} = \frac{N_{o(\text{source})} + N_{o(\text{added})}}{N_{o(\text{source})}} = 1 + \frac{N_{o(\text{added})}}{N_{o(\text{source})}} \tag{2.14}$$

This shows that the minimum possible noise factor, which occurs if the electronics adds no noise, is equal to 1. Noise figure, *NF*, is related to noise factor, *F*, by:

$$\mathrm{NF} = 10\log_{10} F \tag{2.15}$$

Thus, while the noise factor is at least 1, the noise figure is at least 0 dB. In other words, an electronic system that adds no noise has a noise figure of 0 dB.

In the receiver chain, for components with loss (such as switches and filters), the noise figure is equal to the attenuation of the signal. For example, a filter with 3 dB of loss has a noise figure of 3 dB. This is explained by noting that output noise is approximately equal to input noise, but the signal is attenuated by 3 dB. Thus, there has been a degradation of SNR by 3 dB.

Example 2.1: Noise Calculations

Figure 2.2 shows a 50 source resistance loaded with 50 . Determine how much noise voltage per unit bandwidth is present at the output. Also find the noise factor, assuming that R_L does not contribute to noise factor, and compare it to the case where R_L does contribute to noise factor.

Solution:

The noise from the 50 source is $\sqrt{4kTR} = 0.894\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$ at a temperature of 290K, which, after the voltage divider, becomes one half of this value or $v_o = 0.447\mathrm{nV}/\sqrt{\mathrm{Hz}}$.

The complete available power from the source is delivered to the load. In this case;

$$P_o = P_{\text{in (available)}} = kT = 4\times10^{-21}$$

At the output, the complete noise power (available) appears and so, if R_L is noiseless, the noise factor = 1. However, if R_L has noise of $\sqrt{4kTR_L}\ \mathrm{V}/\sqrt{\mathrm{Hz}}$, then at the output, the total noise power is $2kT$ where kT is from R_S and R_L. Therefore, for a resistively matched circuit, the noise figure is 3 dB. Note that the output noise voltage is $0.45\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$ from each resistor for a total of $\sqrt{2}$ $0.45\ \mathrm{nV}/\sqrt{\mathrm{Hz}} = 0.636\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$ (with noise the power adds because the noise voltage is uncorrelated).

Example 2.2: Noise Calculation with Gain Stages

In this example (Figure 2.3), a voltage gain of 20 has been added to the original circuit of Figure 2.2. All resistor values are still 50 . Determine the noise at the

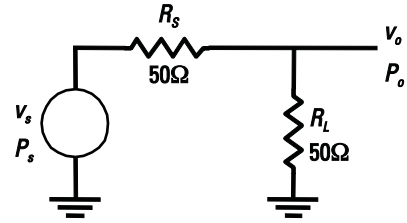

Figure 2.2 Simple circuit used for noise calculations.

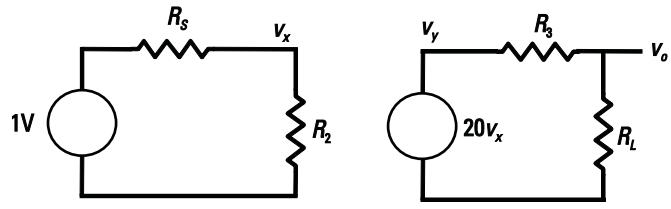

Figure 2.3 Noise calculation with a gain stage.

output of the circuit due to all resistors and then determine the circuit noise figure and signal-to-noise ratio assuming a 1-MHz bandwidth and the input is a 1-V sine wave.

Solution:
In this example, at v_x the noise is still due to only R_S and R_2. As in the previous example, the noise at this point is 0.636 nV/$\sqrt{\text{Hz}}$. The signal at this point is 0.5V; thus, at point v_y, the signal is 10V and the noise due to the two input resistors R_S and R_2 is $0.636 \quad 20 = 12.72$ nV/$\sqrt{\text{Hz}}$. At the output, the signal and noise from the input and output sources see a voltage divider. Thus, one can calculate the individual components. For the combination of R_S and R_2, one obtains:

$$v_{R_s+R_2} = 0.5 \quad 12.72 = 6.36 \text{ nV}/\sqrt{\text{Hz}}$$

The noise from the source can be determined from:

$$v_{R_s} = \frac{6.36 \text{ nV}/\sqrt{\text{Hz}}}{\sqrt{2}} = 4.5 \text{ nV}/\sqrt{\text{Hz}}$$

For the other resistors, the voltage is

$$v_{R_3} = 0.5 \quad 0.9 = 0.45 \text{ nV}/\sqrt{\text{Hz}}$$

$$v_{R_L} = 0.5 \quad 0.9 = 0.45 \text{ nV}/\sqrt{\text{Hz}}$$

Total output noise is given by:

$$v_{\text{no (total)}} = \sqrt{v_{R_3+R_2}^2 + v_{R_3}^2 + v_{R_L}^2} = \sqrt{6.36^2 + 0.45^2 + 0.45^2} = 6.392 \text{ nV}/\sqrt{\text{Hz}}$$

Therefore, the noise figure can now be determined:

$$\text{Noise Factor} = F = \frac{N_{o(\text{total})}}{N_{o(\text{source})}} = \left(\frac{6.392}{4.5}\right)^2 = (1.417)^2 = 2.018$$

$$\text{NF} = 10 \log_{10} F = 10 \log_{10} 2.018 = 3.05 \text{ dB}$$

Since the output voltage also sees a voltage divider of ½, it has a value of 5V. Thus, the signal-to-noise ratio is:

$$\frac{S}{N} = 20\log \frac{5}{\frac{6.392\ \text{nV}}{\sqrt{\text{Hz}}} \times \sqrt{1\ \text{MHz}}} = 117.9\ \text{dB}$$

This example illustrates that noise from the source and amplifier input resistance are the dominant noise sources in the circuit. Each resistor at the input provided 4.5 $\text{nV}/\sqrt{\text{Hz}}$, while the two resistors behind the amplifier each only contribute 0.45 $\text{nV}/\sqrt{\text{Hz}}$. Thus, as explained earlier, after a gain stage, noise is less important.

Example 2.3: Effect of Impedance Mismatch on Noise Figure

Find the noise figure of Example 2.2 again, but now assume that $R_2 = 500$.

Solution:
As before, the output noise due to the resistors is as follows:

$$v_{no\,(R_s)} = 0.9 \quad \frac{500}{550} \quad 20 \quad 0.5 = 8.181\ \text{nV}/\sqrt{\text{Hz}}$$

where 500/550 accounts for the voltage division from the noise source to the node v_x.

$$v_{\text{no}\,(R_2)} = 0.9 \quad \sqrt{10} \quad \frac{50}{550} \quad 20 \quad 0.5 = 2.587\ \text{nV}/\sqrt{\text{Hz}}$$

where the $\sqrt{10}$ accounts for the higher noise in a 500 resistor compared to a 50 resistor.

$$v_{\text{no}\,(R_3)} = 0.9 \quad 0.5 = 0.45\ \text{nV}/\sqrt{\text{Hz}}$$

$$v_{\text{no}\,(R_L)} = 0.9 \quad 0.5\ \text{nV}/\sqrt{\text{Hz}}$$

The total output noise voltage is:

$$v_{\text{no (total)}} = \sqrt{v_{R_s}^2 + v_{R_2}^2 + v_{R_3}^2 + v_{R_L}^2} = \sqrt{8.181^2 + 2.587^2 + 0.45^2 + 0.45^2}$$

$$= 8.604\ \text{nV}/\sqrt{\text{Hz}}$$

$$\text{Noise Factor} = F = \frac{N_{o(\text{total})}}{N_{o(\text{source})}} = \left(\frac{8.604}{8.181}\right)^2 = 1.106$$

$$\text{NF} = 10\log_{10} F = 10\log_{10} 1.106 = 0.438\ \text{dB}$$

Note that this circuit is unmatched at the input. This example illustrates that a mismatched circuit may have better noise performance than a matched one. However, this assumes that it is possible to build a voltage amplifier that requires little power at the input. This may be possible on an IC. However, if transmission lines are included, power transfer will suffer. A matching circuit may need to be added.

2.2.5 The Noise Figure of an Amplifier Circuit

We can now make use of the definition of noise figure just developed and apply it to an amplifier circuit [8]. For the purposes of developing (2.14) into a more useful form, it is assumed that all practical amplifiers can be characterized by an input-referred noise model, such as the one shown in Figure 2.4 where the amplifier is characterized with current gain A_i. (Later chapters will detail how to take a practical amplifier and make it fit this model.) In this model, all noise sources in the circuit are lumped into a series noise voltage source v_n and a parallel current noise source i_{ns} placed in front of a noiseless transfer function.

If the amplifier has finite input impedance, then the input current will be split by some ratio α between the amplifier and the source admittance Y_s:

$$\mathrm{SNR_{in}} = \frac{\alpha^2 {i_{\mathrm{in}}}^2}{\alpha^2 {i_{ns}}^2} \tag{2.16}$$

Assuming that the input referred noise sources are correlated, the output signal-to-noise ratio is

$$\mathrm{SNR_{out}} = \frac{\alpha^2 A_i^2 {i_{\mathrm{in}}}^2}{\alpha^2 A_i^2 ({i_{ns}}^2 + |\, i_n + v_n Y_s\,|^2)} \tag{2.17}$$

Thus, the noise factor can now be written in terms of (2.16) and (2.17):

$$F = \frac{{i_{ns}}^2 + |\, i_n + v_n Y_s\,|^2}{{i_{ns}}^2} = \frac{N_{o(\mathrm{total})}}{N_{o(\mathrm{source})}} \tag{2.18}$$

This can also be interpreted as the ratio of the total output noise to the total output noise due to the source admittance.

In (2.17), it was assumed that the two input noise sources were correlated with each other. In general, they will not be correlated with each other, but rather the current i_n will be partially correlated with v_n and partially uncorrelated. We can expand both the current and voltage into these two explicit parts:

$$i_n = i_c + i_u \tag{2.19}$$

$$v_n = v_c + v_u \tag{2.20}$$

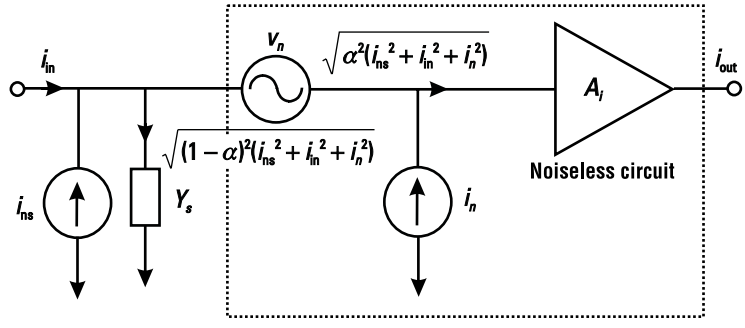

Figure 2.4 Input referred noise model for a device. Current due to v_n is not shown.

In addition, the correlated components will be related by the ratio:

$$i_c = Y_c v_c \tag{2.21}$$

where Y_C is the correlation admittance. Note that we have provided both correlated and uncorrelated terms for both voltage and current, as well as the correlation admittance. These terms are not all independent, making it possible to omit some terms, but we have chosen to keep this section general. Simplifications will be made later when we derive noise factor equations for specific circuits.

The noise factor can now be written as:

$$F = 1 + \frac{i_u{}^2 + |Y_c + Y_s|^2 \; v_c{}^2 + v_u{}^2 \, |Y_s|^2}{i_{ns}{}^2} \tag{2.22}$$

The noise currents and voltages can also be written in terms of equivalent resistance and conductance (these resistors would have the same noise behavior):

$$R_c = \frac{v_c{}^2}{4kT \;\; f} \tag{2.23}$$

$$R_u = \frac{v_u{}^2}{4kT \;\; f} \tag{2.24}$$

$$G_u = \frac{i_u{}^2}{4kT \;\; f} \tag{2.25}$$

$$G_s = \frac{i_{ns}{}^2}{4kT \;\; f} \tag{2.26}$$

Thus, the noise factor is now written in terms of these parameters:

$$F = 1 + \frac{G_u + |Y_c + Y_s|^2 \; R_c + |Y_s|^2 \; R_u}{G_s} \tag{2.27}$$

$$F = 1 + \frac{G_u \;\; \left(G_c + G_s\right)^2 + \left(B_c + B_s\right)^2 \;\; R_c + \;\; G_s{}^2 + B_s{}^2 \;\; R_u}{G_s} \tag{2.28}$$

It can be seen from (2.28) that F is dependent on the equivalent source impedance.

Equation (2.28) can be used not only to determine the noise factor, but also to determine the source loading conditions that will minimize the noise figure. Differentiating with respect to G_s and B_s and setting the derivative to zero, yields the following two conditions for minimum noise (G_{opt} and B_{opt}) after several pages of math:

$$G_{\text{opt}} = \sqrt{\frac{G_u + R_u \;\; \left(\frac{R_c B_c}{R_c + R_u}\right)^2 + G_c{}^2 R_c + \;\; B_c \;\; \left(\frac{R_c B_c}{R_c + R_u}\right)^2 R_c}{R_c + R_u}} \tag{2.29}$$

$$B_{\text{opt}} = \frac{R_c B_c}{R_c + R_u} \tag{2.30}$$

2.2.6 Phase Noise

Radios use reference tones to perform frequency conversion. Ideally, these tones would be perfect and have energy at only the desired frequency. Unfortunately, any real signal source will have energy at other frequencies. Local oscillator noise performance is usually classified in terms of phase noise, which is a measure of how much the output diverges from an ideal impulse function in the frequency domain. We are primarily concerned with noise that causes fluctuations in the phase of the output rather than noise that causes amplitude fluctuations in the tone, since the output typically has a fixed, limited amplitude. The output signal of a reference tone can be described as

$$v_{\text{out}}(t) = V_o \cos(\omega_{\text{LO}} t + \varphi_n(t)) \tag{2.31}$$

Here, $\omega_{\text{LO}}t$ is the desired phase of the output and $\phi_n(t)$ are random fluctuations in the phase of the output due to any one of a number of sources. Phase noise is often quoted in units of dBc/Hz or rad^2/Hz.

The phase fluctuation term $\phi_n(t)$ may be random phase noise or discrete spurious tones, as shown in Figure 2.5. The discrete spurs at a synthesizer output are most likely due to the fractional-N mechanism (discussed in detail in Chapter 10) and the phase noise in an oscillator is mainly due to thermal, flicker, or 1/*f* noise and the finite *Q* of the oscillator tank.

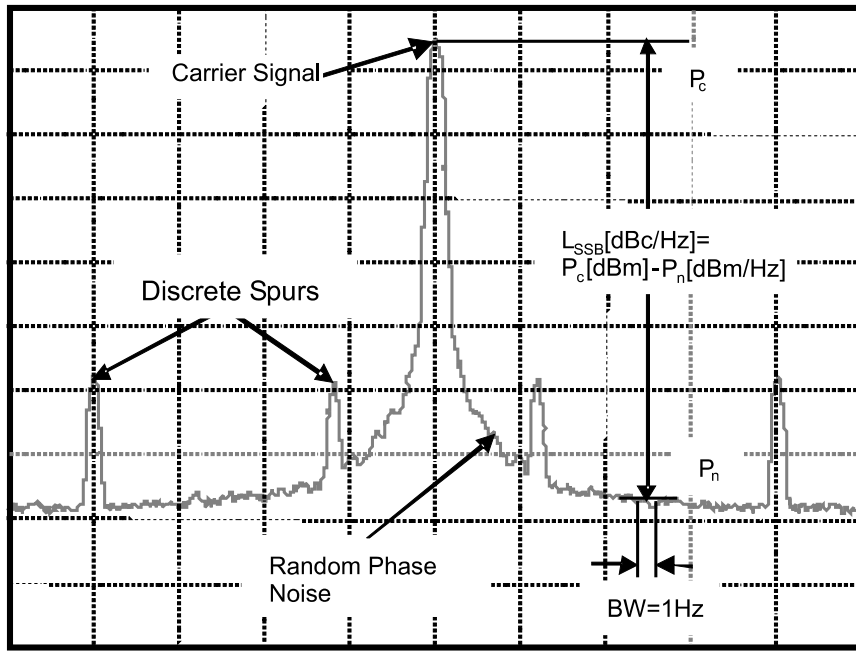

Figure 2.5 An example of phase noise and spurs observed using a spectrum analyzer.

Assume the phase fluctuation is of a sinusoidal form as:

$$\varphi_n(t) = \varphi_p \sin(\omega_m t) \tag{2.32}$$

where ϕ_p is the peak phase fluctuation and ω_m is the offset frequency from the carrier. Substituting (2.32) into (2.31) gives:

$$\begin{aligned} v_{\text{out}}(t) &= V_0 \cos\ \omega_{\text{LO}} t + \varphi_p \sin(\omega_m t) \\ &= V_0\ \cos(\omega_{\text{LO}} t)\cos(\varphi_p \sin(\omega_m t))\quad \sin(\omega_{\text{LO}} t)\sin(\varphi_p \sin(\omega_m t)) \end{aligned} \tag{2.33}$$

For a small phase fluctuation, ϕ_p, (2.33) can be simplified as:

$$\begin{aligned} v_0(t) &= V_0\ \cos(\omega_{\text{LO}} t)\quad \varphi_p \sin(\omega_m t)\sin(\omega_{\text{LO}} t) \\ &= V_0\ \cos(\omega_{\text{LO}} t)\quad \frac{\varphi_p}{2}[\cos(\omega_{\text{LO}}\quad \omega_m)t\quad \cos(\omega_{\text{LO}} + \omega_m)t] \end{aligned} \tag{2.34}$$

It is now evident that the phase-modulated signal includes the carrier signal tone and two symmetric sidebands at an offset frequency, as shown in Figure 2.5. A spectrum analyzer measures the phase-noise power in dBm/Hz, but often phase noise is reported relative to the carrier power as:

$$\varphi_n{}^2(\ \ \omega) = \frac{Noise(\omega_{\text{LO}} +\ \ \omega)}{P_{\text{carrier}}(\omega_{\text{LO}})} \tag{2.35}$$

where *Noise* is the noise power in a 1-Hz bandwidth and P_{carrier} is the power of the carrier or LO tone at the frequency at which the synthesizer is operating. In this form, phase noise has the units of rad^2/Hz. Often this is quoted as so many decibels down from the carrier in dBc/Hz. To further complicate this, both single-sideband and double-sideband phase noise can be defined. Single sideband (SSB) phase noise is defined as the ratio of power in one phase modulation sideband per hertz bandwidth, at an offset ω away from the carrier, to the total signal power. The SSB phase noise power spectral density (PSD) to carrier ratio, in units of [dBc/Hz], is defined as

$$PN_{SSB}(\ \ \omega) = 10\log\ \frac{Noise(\omega_{\text{LO}} +\ \ \omega)}{P_{\text{carrier}}(\omega_{\text{LO}})} \tag{2.36}$$

Combining (2.34) into (2.36) this equation can be rewritten as:

$$PN_{\text{SSB}}(\ \ \omega) = 10\log\ \frac{\frac{1}{2}\ \frac{V_0\varphi_p}{2}^{\ 2}\ \div}{\frac{1}{2}V_0^2} = 10\log\ \frac{\varphi_p^2}{4} = 10\log\ \frac{\varphi_{\text{rms}}^2}{2} \tag{2.37}$$

where ϕ^2_{rms} is the rms phase-noise power density in units of [rad²/Hz]. Note that single-sideband phase noise is by far the most common type reported and often it is not specified as SSB, but rather simply reported as phase noise. However, alternatively, double sideband phase noise is given by:

$$PN_{\text{DSB}}(\quad \omega) = 10\log\ \frac{Noise(\omega_{\text{LO}} + \quad \omega) + Noise(\omega_{\text{LO}} \qquad \omega)}{P_{\text{carrier}}(\omega_{\text{LO}})} = 10\log[\varphi^2_{\text{rms}}] \quad (2.38)$$

From either the single-sideband or double-sideband phase noise, the rms jitter can be obtained as:

$$\varphi_{\text{rms}}(\quad f) = \frac{180}{\pi}\sqrt{10^{\frac{PN_{\text{DSB}}(\quad f)}{10}}} = \frac{180\sqrt{2}}{\pi}\sqrt{10^{\frac{PN_{\text{SSB}}(\quad f)}{10}}}\ [\text{deg}/\sqrt{\text{Hz}}] \quad (2.39)$$

It is also quite common to quote integrated phase noise. The rms integrated phase noise of a synthesizer is given by:

$$\text{IntPN}_{\text{rms}} = \sqrt{\int_{f_1}^{f_2} \varphi_{\text{rms}}{}^2(f)df} \quad (2.40)$$

The limits of integration are usually the offsets corresponding to the lower and upper frequencies of the bandwidth of the information being transmitted.

In addition, it should be noted that dividing or multiplying a signal in the time domain also multiplies or divides the phase noise. Thus, if a signal frequency is multiplied by N, then the phase noise is related by:

$$\begin{aligned} \varphi^2{}_{\text{rms}}(N\omega_{\text{LO}} + \quad \omega) &= N^2 \times \varphi^2{}_{\text{rms}}(\omega_{\text{LO}} + \quad \omega) \\ \varphi^2{}_{\text{rms}}\ \ \frac{\omega_{\text{LO}}}{N} + \quad \omega_{\div} &= \frac{\varphi^2{}_{\text{rms}}(\omega_{\text{LO}} + \quad \omega)}{N^2} \end{aligned} \quad (2.41)$$

Note this assumes that the circuit that did the frequency translation is noiseless. Also, note that the phase noise is scaled by N^2 rather than N to get units of $\mathbf{V}^2$ rather than noise voltage.

2.3 Linearity and Distortion in RF Circuits

In an ideal system, the output is linearly related to the input. However, in any real device the transfer function is usually a lot more complicated. This can be due to active or passive devices in the circuit, or the signal swing being limited by the power supply rails. Unavoidably, the gain curve for any component is never a perfectly straight line, as illustrated in Figure 2.6.

The resulting waveforms can appear as shown in Figure 2.7. For amplifier saturation, typically the top and bottom portions of the waveform are clipped equally,

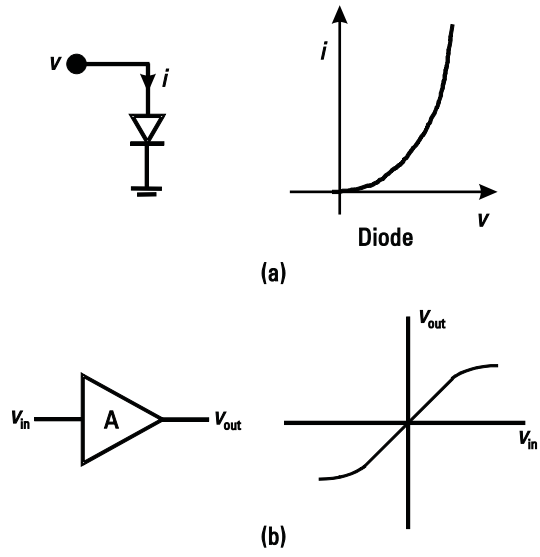

Figure 2.6 Illustration of the nonlinearity in (a) a diode and (b) an amplifier.

as shown in Figure 2.7(b). However, if the circuit is not biased between the two clipping levels, then clipping can be nonsymmetrical as shown in Figure 2.7(c).

2.3.1 Power Series Expansion

Mathematically, any nonlinear transfer function can be written as a series expansion of power terms unless the system contains memory, in which case a Volterra series is required [9, 10]:

$$v_{\text{out}} = k_0 + k_1 v_{\text{in}} + k_2 {v_{\text{in}}}^2 + k_3 {v_{\text{in}}}^3 + \ldots \tag{2.42}$$

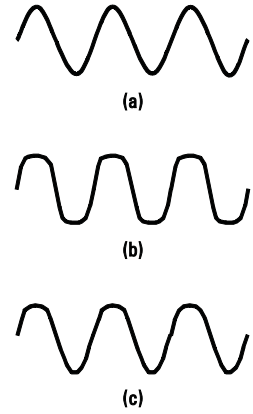

Figure 2.7 Distorted output waveforms: (a) input; (b) output, clipping; and (c) output, bias wrong.

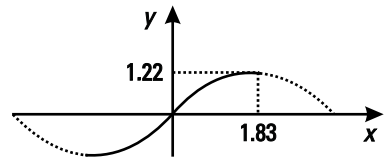

Figure 2.8 An example of a nonlinearity with first- and third-order terms.

To describe the nonlinearity perfectly, an infinite number of terms is required; however, in many practical circuits the first three terms are sufficient to characterize the circuit with a fair degree of accuracy.

Symmetrical saturation as shown in Figure 2.6(b) can be modeled with odd order terms; for example,

$$y = x \quad \frac{1}{10}x^3 \tag{2.43}$$

looks like Figure 2.8. Another example is an exponential nonlinearity as shown in Figure 2.6(a), which has the form

$$x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \tag{2.44}$$

This contains both even and odd power terms, because it does not have symmetry about the y-axis. Real circuits will have more complex power series expansions.

One common way of characterizing the linearity of a circuit is called the two-tone test. In this test, an input consisting of two sine waves is applied to the circuit.

$$v_{\text{in}} = v_1 \cos\omega_1 t + v_2 \cos\omega_2 t = X_1 + X_2 \tag{2.45}$$

When this tone is applied to the transfer function given in (2.42), the result is a number of terms:

$$v_0 = k_0 + \underbrace{k_1(X_1 + X_2)}_{\text{desired}} + \underbrace{k_2(X_1 + X_2)^2}_{\text{second order}} + \underbrace{k_3(X_1 + X_2)^3}_{\text{third order}} \tag{2.46}$$

$$v_0 = k_0 + k_1(X_1 + X_2) + k_2(X_1^2 + 2X_1X_2 + X_2^2) + k_3(X_1^3 + 3X_1^2X_2 + 3X_1X_2^2 + X_2^3) \tag{2.47}$$

These terms can be further broken down into various frequency components. For instance, the X_1^2 term has a component at dc and another at the second harmonic of the input:

$$X_1^2 = (v_1 \cos\omega_1 t)^2 = \frac{v_1^2}{2}(1 + \cos 2\omega_1 t) \tag{2.48}$$

The second-order terms can be expanded as follows:

$$(X_1 + X_2)^2 = \underbrace{X_1^2}_{\substack{\text{dc+}\\ \text{HD2}}} + \underbrace{2X_1X_2}_{\text{MIX}} + \underbrace{X_2^2}_{\substack{\text{dc+}\\ \text{HD2}}} \tag{2.49}$$

where second-order terms are comprised of second harmonics HD2, and mixing components, here labeled MIX, but sometimes labeled IM2 for second-order intermodulation. The mixing components will appear at the sum and difference frequencies of the two input signals. Note also that second-order terms cause an additional dc term to appear.

The third-order terms can be expanded as follows:

$$(X_1 + X_2)^3 = \underbrace{X_1^3}_{\substack{\text{FUND}\\ +\text{HD3}}} + \underbrace{3X_1^2X_2}_{\substack{\text{IM3}+\\ \text{FUND}}} + \underbrace{3X_1X_2^2}_{\substack{\text{IM3}+\\ \text{FUND}}} + \underbrace{X_2^3}_{\substack{\text{FUND}\\ +\text{HD3}}} \tag{2.50}$$

Third-order nonlinearity results in third-order harmonics HD3 and third-order intermodulation IM3. Expansion of both the HD3 terms and the IM3 terms show output signals appearing at the input frequencies. The effect is that third-order nonlinearity can change the gain, which is seen as gain compression. This is summarized in Table 2.1.

Note that in the case of an amplifier, only the terms at the input frequency are desired. Of all the unwanted terms, the last two at frequencies $2\omega_1 \quad \omega_2$ and $2\omega_2 \quad \omega_1$ are the most troublesome, since they can fall in the band of the desired outputs if ω_1 is close in frequency to ω_2 and therefore cannot be easily filtered out. These two tones are usually referred to as third-order intermodulation terms (IM3 products).

Table 2.1 Summary of Distortion Components

Frequency	*Component Amplitude*
dc	$k_o + \frac{k_2}{2}(v_1^2 + v_2^2)$
ω_1	$k_1v_1 + k_3v_1 \quad \frac{3}{4}v_1^2 + \frac{3}{2}v_2^2 \div$
ω_2	$k_1v_2 + k_3v_2 \quad \frac{3}{4}v_2^2 + \frac{3}{2}v_1^2 \div$
$2\omega_1$	$\frac{k_2v_1^2}{2}$
$2\omega_2$	$\frac{k_2v_2^2}{2}$
$\omega_1 \pm \omega_2$	$k_2v_1v_2$
$\omega_2 \pm \omega_1$	$k_2v_1v_2$
$3\omega_1$	$\frac{k_3v_1^3}{4}$
$3\omega_2$	$\frac{k_3v_2^3}{4}$
$2\omega_1 \quad \omega_2$	$\frac{3}{4}k_3v_1^2v_2$
$2\omega_2 \quad \omega_1$	$\frac{3}{4}k_3v_1v_2^2$

Table 2.2 Outputs from Nonlinear Circuits with Inputs at $f_1 = 7$, $f_2 = 8$ MHz

	Symbolic Frequency	*Example Frequency*	*Name*	*Comment*
First order	f_1, f_2	7, 8	Fundamental	Desired output
Second order	$2f_1, 2f_2$	14, 16	HD2 (Harmonics)	Can filter
	$f_2 - f_1, f_2 + f_1$	1, 15	IM2 (Mixing)	Can filter
Third order	$3f_1, 3f_2$	21, 24	HD3 (Harmonic)	Can filter harmonics
	$2f_1 - f_2$,	6	IM3 (Intermod)	Close to fundamental,
	$2f_2 - f_1$	9	IM3 (Intermod)	difficult to filter

Example 2.4: Determination of Frequency Components Generated in a Nonlinear System

Consider a nonlinear circuit with 7-MHz and 8-MHz tones applied at the input. Determine all output frequency components, assuming distortion components up to the third order.

Solution:
Table 2.2 and Figure 2.9 show the outputs.

It is apparent that harmonics can be filtered out easily, while the third-order intermodulation terms, being close to the desired tones, may be difficult to filter.

2.3.2 Third-Order Intercept Point

One of the most common ways to test the linearity of a circuit is to apply two signals at the input, having equal amplitude and offset by some frequency, and plot fundamental output and intermodulation output power as a function of input power as shown in Figure 2.10. From the plot, the third-order intercept point (IP3) is determined. The third-order intercept point is a theoretical point where the amplitudes of the intermodulation tones at $2\omega_1 \quad \omega_2$ and $2\omega_2 \quad \omega_1$ are equal to the amplitudes of the fundamental tones at ω_1 and ω_2.

From Table 2.1, if $v_1 = v_2 = v_i$, then the fundamental is given by

$$\text{fund} = k_1 v_i + \frac{9}{4} k_3 v_i^3 \tag{2.51}$$

The linear component of (2.51) given by:

$$\text{fund} = k_1 v_i \tag{2.52}$$

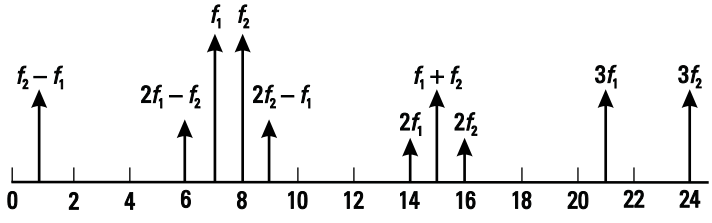

Figure 2.9 The output spectrum with inputs at 7 and 8 MHz.

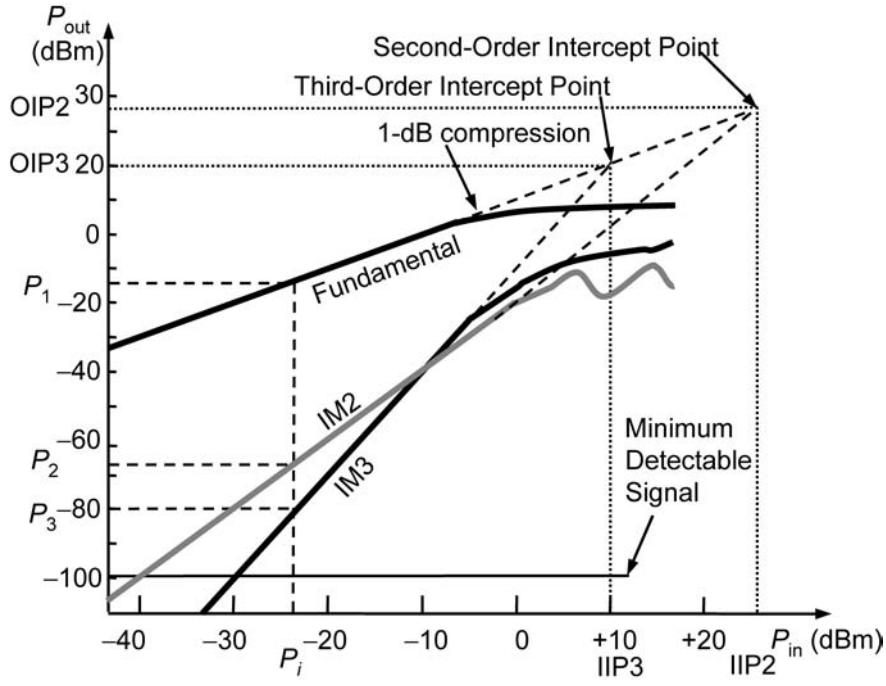

Figure 2.10 Plot of output power of fundamental, IM3, and IM2 versus input power.

can be compared to the third-order intermodulation term given by:

$$\text{IM3} = \frac{3}{4} k_3 v_i^3 \tag{2.53}$$

Note that for small v_i, the fundamental rises linearly (20 dB/decade) and the IM3 terms rise as the cube of the input (60 dB/decade). A theoretical voltage at which these two tones will be equal can be defined:

$$\frac{\frac{3}{4} k_3 {v_{\text{IP3}}}^3}{k_1 v_{\text{IP3}}} = 1 \tag{2.54}$$

This can be solved for v_{IP3}:

$$v_{\text{IP}\,3} = 2\sqrt{\frac{k_1}{3|k_3|}} \tag{2.55}$$

Note that for a circuit experiencing compression, k_3 must be negative. Thus in (2.66) the absolute value of k_3 must be used. Note also that (2.55) gives the input voltage at the third-order intercept point. The input power at this point is called the input third-order intercept point (IIP3). If IP3 is specified at the output, it is called the output third-order intercept point (OIP3).

Of course, the third-order intercept point cannot actually be measured directly, since by the time the amplifier reached this point, it would be heavily overloaded. Therefore, it is useful to describe a quick way to extrapolate it at a given power level. Assume that a device with power gain G has been measured to have an output power of P_1 at the fundamental frequency and a power of P_3 at the IM3 frequency for a given input power of P_i, as illustrated in Figure 2.10. Now, on a log plot (for example when power is in dBm) of P_3 and P_1 versus P_i, the IM3 terms have a slope of 3 and the fundamental terms have a slope of 1. Therefore:

$$\frac{\text{OIP3} \quad P_1}{\text{IIP3} \quad P_i} = 1 \tag{2.56}$$

$$\frac{\text{OIP3} \quad P_3}{\text{IIP3} \quad P_i} = 3 \tag{2.57}$$

since subtraction on a log scale amounts to division of power.

Also noting that:

$$G = \text{OIP3} \quad \text{IIP3} = P_1 \quad P_i \tag{2.58}$$

These equations can be solved to give:

$$\text{IIP3} = P_1 + \frac{1}{2}[P_1 \quad P_3] \quad G = P_i + \frac{1}{2}[P_1 \quad P_3] \tag{2.59}$$

2.3.3 Second-Order Intercept Point

A second-order intercept point (IP2) can be defined similarly to the third-order intercept point. Which one is used depends largely on which is more important in the system of interest; for example, second-order distortion is particularly important in direct down-conversion receivers.

If two tones are present at the input, then the second-order output is given by

$$v_{\text{IM2}} = k_2 v_i^2 \tag{2.60}$$

Note that the IM2 terms rise at 40 dB/dec rather than at 60 dB/dec, as in the case of the IM3 terms as shown in Figure 2.10.

The theoretical voltage at which the IM2 term will be equal to the fundamental term given in (2.52) can be defined:

$$\frac{k_2 v_{\text{IP2}}^2}{k_1 v_{\text{IP2}}} = 1 \tag{2.61}$$

This can be solved for v_{IP2}:

$$v_{\text{IP2}} = \frac{k_1}{k_2} \tag{2.62}$$

Assume that a device with power gain G has been measured to have an output power of P_1 at the fundamental frequency and a power of P_2 at the IM2 frequency for a given input power of P_i, as illustrated in Figure 2.10. Now, on a log plot (for example when the power is in dBm) of P_2 and P_1 versus P_i, the IM2 terms have a slope of 2 and the fundamental terms have a slope of 1. Therefore:

$$\frac{\text{OIP2} \quad P_1}{\text{IIP2} \quad P_i} = 1 \tag{2.63}$$

$$\frac{\text{OIP2} \quad P_2}{\text{IIP2} \quad P_i} = 2 \tag{2.64}$$

since subtraction on a log scale amounts to division of power.

Note that

$$G = \text{OIP2} \quad \text{IIP2} = P_1 \quad P_i \tag{2.65}$$

These equations can be solved to give:

$$\text{IIP2} = P_1 + [P_1 \quad P_2] \quad G = P_i + [P_1 \quad P_2] \tag{2.66}$$

2.3.4 The 1-dB Compression Point

In addition to measuring the IP3 or IP2 of a circuit, the 1-dB compression point is another common way to measure linearity. This point is more directly measurable than IP3 and requires only one tone rather than two (although any number of tones can be used). The 1-dB compression point is simply the power level, specified at either the input or the output, where the output power is 1 dB less than it would have been in an ideally linear device. It is also marked in Figure 2.10.

We first note that at 1-dB compression, the ratio of the actual output voltage v_o to the ideal output voltage v_{oi} is

$$20\log_{10} \quad \frac{v_o}{v_{oi}} \div = \quad 1\,\text{dB} \tag{2.67}$$

or

$$\frac{v_o}{v_{oi}} = 0.89125 \tag{2.68}$$

Now referring again to Table 2.1, we note that the actual output voltage for a single tone is

$$v_o = k_1 v_i + \frac{3}{4} k_3 {v_i}^3 \tag{2.69}$$

for an input voltage v_i. The ideal output voltage is given by

$$v_{\text{oi}} = k_1 v_i \tag{2.70}$$

Thus, the 1-dB compression point can be found by substituting (2.69) and (2.70) into (2.68):

$$\frac{k_1 v_{1\text{dB}} + \frac{3}{4} k_3 {v_{1\text{dB}}}^3}{k_1 v_{1\text{dB}}} = 0.89125 \tag{2.71}$$

Note that for a nonlinearity that causes compression, rather than one that causes expansion, k_3 has to be negative. Solving (2.71) for $v_{1\text{dB}}$ gives:

$$v_{1\text{dB}} = 0.38\sqrt{\frac{k_1}{|k_3|}} \tag{2.72}$$

If more than one tone is applied, the 1-dB compression point will occur for a lower input voltage. In the case of two equal amplitude tones applied to the system, the actual output power for one frequency is:

$$v_o = k_1 v_i + \frac{9}{4} k_3 {v_i}^3 \tag{2.73}$$

The ideal output voltage is still given by (2.70). So now the ratio is

$$\frac{k_1 v_{1\text{dB}} + \frac{9}{4} k_3 {v_{1\text{dB}}}^3}{k_1 v_{1\text{dB}}} = 0.89125 \tag{2.74}$$

Therefore, the 1-dB compression voltage is now:

$$v_{1\text{dB}} = 0.22\sqrt{\frac{k_1}{|k_3|}} \tag{2.75}$$

Thus, as more tones are added, this voltage will continue to get lower.

2.3.5 Relationships Between 1-dB Compression and IP3 Points

In the last two sections, formulas for the IP3 and the 1-dB compression points have been derived. Since we now have expressions for both these values, we can find a relationship between these two points. Taking the ratio of (2.55) and (2.72) gives

$$\frac{v_{\text{IP3}}}{v_{1\text{dB}}} = \frac{2\sqrt{\frac{k_1}{3|k_3|}}}{0.38\sqrt{\frac{k_1}{|k_3|}}} = 3.04 \tag{2.76}$$

Thus, these voltages are related by a factor of 3.04 or about 9.66 dB, independent of the particulars of the nonlinearity in question. In the case of the 1-dB compression point with two tones applied, the ratio is larger. In this case:

$$\frac{v_{IP3}}{v_{1dB}} = \frac{2\sqrt{\frac{k_1}{3|k_3|}}}{0.22\sqrt{\frac{k_1}{|k_3|}}} = 5.25 \tag{2.77}$$

Thus, these voltages are related by a factor of 5.25 or about 14.4 dB.

Thus, one can estimate that for a single tone, the compression point is about 10 dB below the intercept point, while for two tones, the 1-dB compression point is close to 15 dB below the intercept point. The difference between these two numbers is just the factor of three (4.77 dB) resulting from the second tone.

Note that this analysis is valid for third-order nonlinearity. For stronger nonlinearity (i.e., containing fifth-order terms), additional components are found at the fundamental as well as at the intermodulation frequencies. Nevertheless, the above is a good estimate of performance.

Example 2.5: Determining IIP3 and 1-dB Compression Point from Measurement Data

An amplifier designed to operate at 2 GHz with a gain of 10 dB has two signals of equal power applied at the input. One is at a frequency of 2.0 GHz and another at a frequency of 2.01 GHz. At the output, four tones are observed at 1.98, 2.0, 2.01, and 2.02 GHz. The power levels of the tones are –70, –20, –20, and –70 dBm, respectively. Determine the IIP3 and 1-dB compression point for this amplifier.

Solution:
The tones at 1.98 and 2.02 GHz are the IP3 tones. We can use (2.59) directly to find the IIP3

$$\text{IIP3} = P_1 + \frac{1}{2}[P_1 \quad P_3] \quad G = \quad 20 + \frac{1}{2}[\ \ 20 + 70] \quad 10 = \quad 5 \text{ dBm}$$

The 1-dB compression point for a single tone is 9.66 dB lower than this value, about –14.7 dBm at the input.

2.3.6 Broadband Measures of Linearity

Intercept points and 1-dB compression points are two common measures of linearity, but they are by no means the only ones. Two other measures of linearity that are common in wideband systems, which handle many signals simultaneously are called composite triple-order beat (CTB) and composite second-order beat (CSO) [11, 12]. In these tests of linearity, N signals of voltage v_i are applied to the circuit equally spaced in frequency, as shown in Figure 2.11. Note here that, as an example, the tones are spaced 6 MHz apart (this is the spacing for a cable television system, for which this is a popular way to characterize linearity). Note also that the tones are never placed at a frequency that is an exact multiple of the spacing (in this case 6 MHz). This is done so that third-order terms and second-order terms fall at different frequencies. This will be clarified shortly.

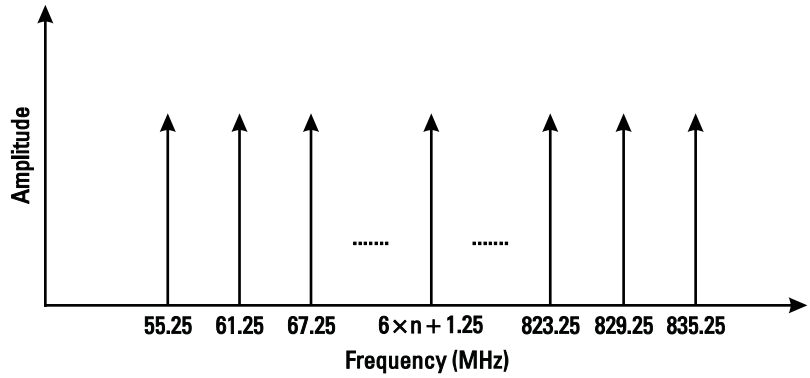

Figure 2.11 Equally spaced tones entering a broadband circuit.

If we take three of these signals, then the third-order nonlinearity gets a little more complicated than before:

$$(x_1 + x_2 + x_3)^3 = \underbrace{x_1^3 + x_2^3 + x_3^3}_{\text{HD3}} + \underbrace{3x_1^2x_2 + 3x_1^2x_3 + 3x_2^2x_1 + 3x_3^2x_1 + 3x_2^2x_3 + 3x_3^2x_2}_{\text{IM3}} + \underbrace{6x_1x_2x_3}_{\text{TB}} \tag{2.78}$$

The last term in the expression causes CTB, in that it creates terms at frequencies $\omega_1 \pm \omega_2 \pm \omega_3$ of magnitude $1.5k_3v_i$ where $\omega_1 < \omega_2 < \omega_3$. This is twice as large as the IM3 products. Note that, except for the case where all three are added ($\omega_1 + \omega_2 + \omega_3$), these tones can fall into any of the channels being used and many will fall into the same channel. For instance, in Figure 2.11, 67.25 73.25 + 79.25 = 73.25 MHz, or 49.25 55.25 + 79.25 = 73.25 MHz will both fall on the 73.25-MHz frequency. In fact, there will be many more triple beat (TB) products than IM3 products. Thus, these terms become more important in a wide band system. It can be shown that the maximum number of terms will fall on the tone at the middle of the band. With N tones, it can be shown that the number of tones falling there will be

$$\text{Tones} = \frac{3}{8}N^2 \tag{2.79}$$

We have already said that the voltage of these tones is twice that of the IM3 tones. We also note here that if the signal power is backed off from the IP3 power by some amount, then the power in the IM3 tones will be backed off three times as much (calculated on a logarithmic scale). Therefore, if each fundamental tone is at a power level of P_s, then the power of the TB tones will be:

$$\text{TB(dBm)} = P_{\text{IP3}} \quad 3(P_{\text{IP3}} \quad P_s) + 6 \tag{2.80}$$

where P_{IP3} is the IP3 power level for the given circuit.

Now, assuming that all tones add as power rather than voltage, and noting that CTB is usually specified as so many decibels down from the signal power,

$$\text{CTB(dB)} = P_s \quad P_{\text{IP3}} \quad 3(P_{\text{IP3}} \quad P_s) + 6 + 10\log\left[\frac{3}{8}N^2\right] \tag{2.81}$$

Note that CTB can be found using either input or output referred power levels.

The CSO is similar to the CTB, and can be used to measure the linearity of a broadband system. Again, if we have N signals all at the same power level, we now consider the second-order distortion products of each pair of signals that falls at frequencies $\omega_1 \pm \omega_2$. In this case, the signals fall at frequencies either above or below the carriers rather than right on top of them unlike the triple beat terms, provided that the carriers are not some even multiple of the channel spacing. For example, in Figure 2.11, 49.25 + 55.25 = 104.5 MHz. This is 1.25 MHz above the closest carrier at 103.25 MHz. All the sum terms will fall 1.25 MHz above the closest carrier, while the difference terms such as 763.25 – 841.25 = 78 will fall 1.25 MHz below the closest carrier at 79.25 MHz. Thus, the second-order and third-order terms can be measured separately. The number of terms that fall next to any given carrier will vary. Some of the $\omega_1 + \omega_2$ terms will fall out of band and the maximum number in band will fall next to the highest frequency carrier. The number of second-order beats above any given carrier is given by

$$N_B = (N \quad 1)\frac{f \quad 2f_L + d}{2(f_H \quad f_L)} \tag{2.82}$$

where N is the number of carriers, f is the frequency of the measurement channel, f_L is the frequency of the lowest channel, f_H is the frequency of the highest channel, and d is the frequency offset from a multiple of the channel spacing (1.25 MHz in Figure 2.11).

For the case of the difference frequency second-order beats, there are more of these at lower frequencies, and the maximum number will be next to the lowest frequency carrier. In this case, the number of second-order products next to any carrier can be approximated by:

$$N_B = (N \quad 1) \quad 1 \quad \frac{f \quad d}{f_H \quad f_L} \div \tag{2.83}$$

Each of the second-order beats is an IP2 tone. Therefore, if each fundamental tone is at a power level of P_s, then the power of the second-order beat (SO) tones will be:

$$\mathrm{SO(dBm)} = P_{\mathrm{IP2}} \quad 2(P_{\mathrm{IP2}} \quad P_s) \tag{2.84}$$

Thus, the composite second-order beat product will be given by:

$$\mathrm{CSO(dB)} = P_s \quad [P_{\mathrm{IP2}} \quad 2(P_{\mathrm{IP2}} \quad P_s) + 10\log(N_B)] \tag{2.85}$$

2.4 Modulated Signals

Radio frequency transceivers are required because it is not feasible to build an antenna that will transmit signals at frequencies close to dc. So far in this chapter only sinusoidal tones have been discussed, but in order for these tones to convey any useful information we must change one or more of their properties to convey informa-

tion across the link. This process is called modulating the carrier. Usually in radio frequency communication systems, the bandwidth that the data occupies is a small fraction of the frequency at which it is transmitted. There are many types of modulation (ways of encoding data onto a carrier for transmission) in use today. In some applications, the data rate is low and therefore a simple modulation scheme is adequate and preferred as it reduces the requirements on the radio. Other applications require a fast data rate and therefore a much higher performance radio. In this section, only digital modulations will be discussed. While certain legacy products may still send analog information over a link, most modern radios first convert the information into a stream of bits, which are then transmitted. Once the data is received at the other end of the link and demodulated, digital signal processors are then responsible for converting this data back into voice, video, or whatever was being transmitted.

In the simplest case, at low data rates one bit at a time may be transmitted in a digital communication link. In more complex systems, multiple bits may be transmitted simultaneously by using symbols that represent more than one bit. In general, the data stream is random and therefore it has power over a range of frequencies. By comparison, a pure square wave has power at one frequency and at odd harmonics of that frequency. The power spectral density of a data stream of symbol rate T_S at baseband can be approximated as:

$$\mathrm{PSD}_{\mathrm{BB}}(f) = A \times T_S^2 \frac{\sin^2(\pi f T_S)}{(\pi f T_S)^2} \tag{2.86}$$

where A is a constant of proportionality. Equation (2.86) is plotted in Figure 2.12. Note that most of the power in the signal is concentrated in frequencies below the symbol frequency $f_S = 1/T_S$. In theory you can recover the information if you limit the bandwidth of the transmission to $0.5f_s$ (also known as the Nyquist bandwidth, which is the minimum bandwidth to avoid intersymbol interference), but in practice it is common to limit the spectrum of the transmitted information to a somewhat higher frequency. If we assume that the signal is limited to f_S, then the transmission efficiency is one symbol per second per hertz of bandwidth used (1 symbol/sec/Hz). In order to transmit more data in a given bandwidth, symbols that represent more bits are used. Thus, if symbols are used that represent two bits, the transmission efficiency would be roughly 2 bits/sec/Hz, and if a symbol represents four bits then the transmission efficiency would be 4 bits/sec/Hz and so on.

Now let us suppose that a radio signal has a bandwidth of BW and a total rms noise power of N is present in the data bandwidth. The power spectral density (PSD) of the noise is given by

$$N_o = \frac{N}{BW} \tag{2.87}$$

Now the energy per symbol, E_s, is the average signal power S multiplied by the time period T_s over which the symbol is transmitted

$$E_s = S \times T_s \tag{2.88}$$

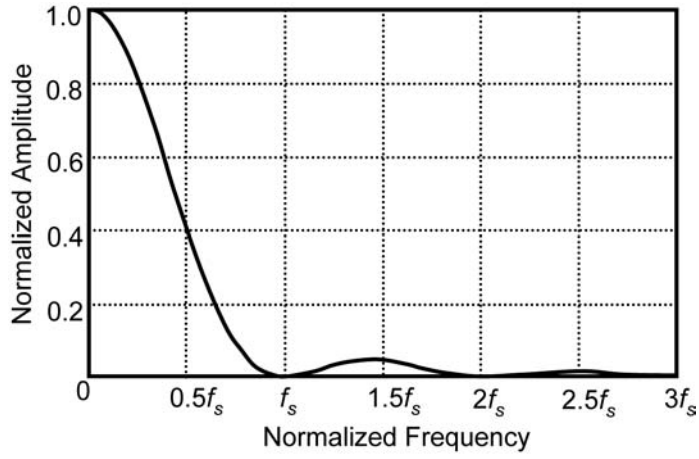

Figure 2.12 Power spectral density of a data stream at baseband.

Thus, the ratio of E_s to the PSD of the noise N_o is given by

$$\frac{E_s}{N_o} = \frac{S \times T_s}{\frac{N}{BW}} = \frac{S}{N} \times \frac{BW}{f_s} \tag{2.89}$$

Thus, the ratio of the energy per symbol to the noise density is equal to the signal-to-noise ratio of the radio if the radio channel bandwidth is equal to the symbol frequency. This result is important because it will allow us to relate concepts in digital modulation to the SNR requirement of the radio.

It is important to note that in digital communications it is often more common to see E_b, the energy per bit even when symbols are used. E_b is related to E_s by:

$$E_b = \frac{E_s}{\log_2 M} \tag{2.90}$$

where $\log_2 M$ is the number of bits transmitted per symbol. Equation (2.89) can be rewritten as:

$$\frac{E_b}{N_o} = \frac{S}{N} \times \frac{BW}{f_s} \times \frac{1}{\log_2 M} \tag{2.91}$$

Having discussed data and the concept of symbols, the next sections will describe various modulation types.

2.4.1 Phase Modulation

In order to transmit information, more than just a carrier tone is needed. The tone has to change in some way over time to indicate what information is being sent. In general, a carrier has two properties that can be changed or modulated in order to convey information: amplitude and phase. We note that frequency modulation is

not treated separately as it can be understood to be a subset of phase modulation. Some modulation schemes only change the amplitude of the signal and some only change the phase (or frequency) of the signal, but some more complicated modulation schemes change both. We will begin by considering modulation that changes only the phase of the carrier.

2.4.1.1 Phase Shift Keying (PSK)

A phase shift keyed (PSK) modulated signal encodes data by changing the phase of the carrier signal according to which bits are to be transmitted. A PSK signal is given by:

$$S_{\mathrm{PSK}}(t) = A \times \cos(\omega_{\mathrm{RF}} t + \phi_{\mathrm{bits}}) \tag{2.92}$$

where A is the amplitude of the carrier signal (a constant), ω_{RF} is the frequency of the carrier, and ϕ_{bits} is the phase, which is given by:

$$\begin{gathered} \phi_{\mathrm{bits}} = \frac{2\pi}{2^M} \times (i \quad 1) \\ i = 1,2,3 \ldots M \end{gathered} \tag{2.93}$$

where $\log_2 M$ is the number of bits transmitted per phase change. For instance, if only one bit is transmitted per phase change (called binary phase shift keying or BPSK) then the phase is either 0° or 180° (i = 0 or 1) depending on whether a zero or a one is transmitted. This is illustrated in the time domain in Figure 2.13. If two bits are transmitted per phase change (called quadrature phase shift keying or QPSK) then four phases are required. Similarly, three bits can be transmitted if eight phases are employed and so on. So, why not use an infinite number of phases to get an infinite bit rate with this modulation scheme? The answer is that in the presence of noise, the more phases that are used (and therefore the closer adjacent phases are to one another), the harder it is to determine one from another; this increases the probability that an error will occur.

In general, the phase angle in PSK modulation can be represented as a vector consisting of in phase (I) and quadrature (Q) components as shown in Figure 2.14.

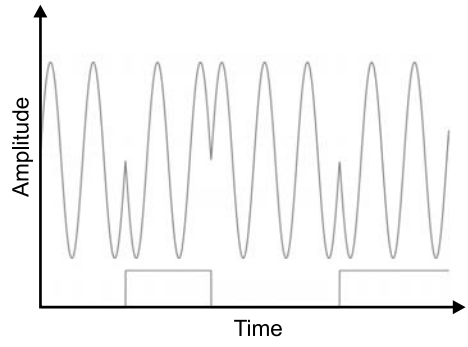

Figure 2.13 Example of a BPSK waveform.

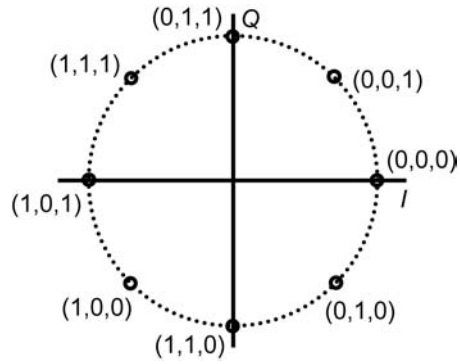

Figure 2.14 Phase plot of an 8-PSK modulated signal.

A very important feature in this figure is that adjacent symbols differ by only one bit. This means that if a phase is misinterpreted, a minimum number of bit errors will occur.

One additional refinement that is sometimes implemented with QPSK is to delay one of the baseband date streams (usually the *Q* path) by half a symbol period. This means that the *I* and *Q* data never change at the same time. Thus, instead of getting a maximum possible instantaneous phase shift of 180°, the maximum instantaneous phase shift is only 90°. This makes the modulation more spectrally efficient. When this is done, the modulation is often referred to as offset quadrature phase shift keying (OQPSK).

Beyond determining the basic functional parts of different PSK transceivers, it is necessary to determine the required SNR for a given probability of a symbol error. This allows the completion of the system specifications. As a starting point, a given phase is transmitted, and in the course of that transmission, noise is added that will tend to change the phase of the transmitted signal. If the phase is changed enough, it will pass a threshold and the symbol will be incorrectly interpreted, resulting in a symbol error. This is illustrated in Figure 2.15 [12]. Here, the nominal transmitted phase is at 0° and is at the center of a normal probability density function. The threshold for misinterpreting a symbol is $\pm\pi/M$ (the shaded area indicates when an

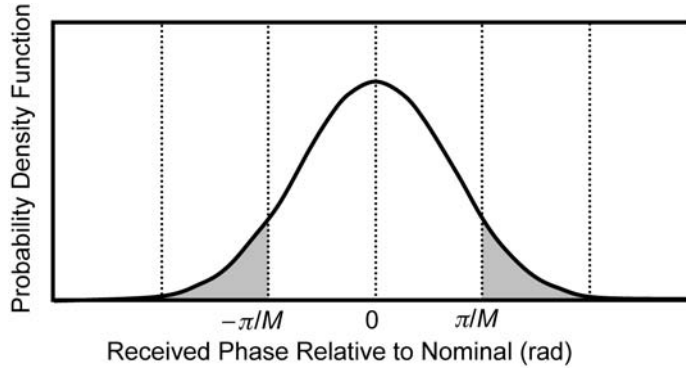

Figure 2.15 Probability density function of a PSK signal showing the probability of making a signal error.

error has been made). If the noise changes the phase by more than this amount, the symbol will be incorrectly interpreted. Note that as long as only one bit changes when an adjacent phase is incorrectly interpreted, the bit error rate as a function of the symbol error rate is given by:

$$P_b \quad \frac{P_s}{\log_2 M} \tag{2.94}$$

A plot of the probability of symbol error versus E_b/N_o can be derived based on statistics and the above discussion. An approximate formula for the probability of symbol error as a function of E_b/N_o is given by [13]:

$$P_s \quad M, \frac{E_b}{N_o} \div \quad 2Q \quad \sqrt{2 \frac{E_b}{N_o} \log_2(M)} \times \sin \quad \frac{\pi}{M} \div \div \tag{2.95}$$

where $Q(x)$ is the area underneath the tail of a Gaussian probability density function. $Q(x)$ is given by:

$$Q(x) = \quad \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}} dz \tag{2.96}$$

Equation (2.95) can also be easily converted to give the probability of bit error using (2.94):

$$P_B \quad M, \frac{E_s}{N_o} \div \quad \frac{2}{\log_2 M} \times Q \quad \sqrt{2 \frac{E_s}{N_o}} \times \sin \quad \frac{\pi}{M} \div \div \tag{2.97}$$

The probability of bit error is plotted in Figure 2.16. Obviously, as the number of symbols increases at a given E_s/N_o, the probability of error is larger.

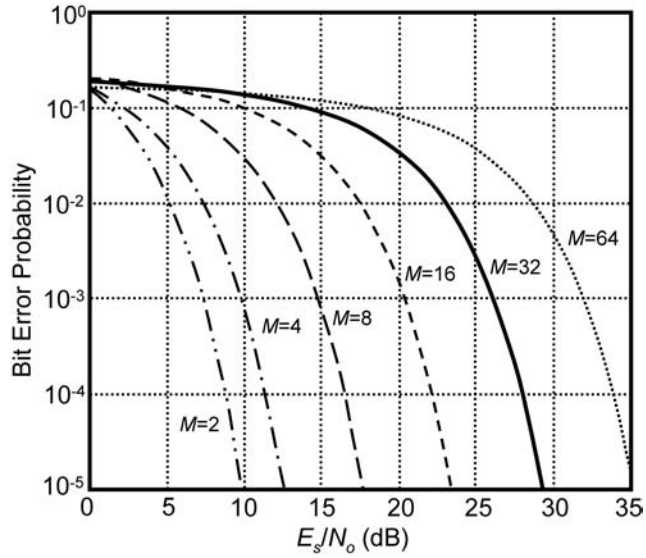

Figure 2.16 Probability of bit error versus E_s/N_o for PSK.

Example 2.6: The Required Noise Figure of a Radio

A receiver must be able to detect a signal at a power level of 95 dBm with a bit error rate of 10^{-3}. The channel is 1 MHz wide at baseband and two modulation formats are used. One is BPSK and the other is 8-PSK. Determine the required receiver NF in both cases.

Solution:
A bandwidth of 1 MHz means that a data rate of 1 megasymbol per second is possible assuming for simplicity that the bandwidth is equal to the symbol rate. For BPSK, the data rate will be 1 Mbps, and for 8-PSK there are 3 bits per symbol so this link will be able to transmit 3 Mbps of information. From Figure 2.16, to achieve an error rate of 10^{-3} for BPSK we need an E_s/N_o of about 7.3 dB, and for 8-PSK a bit error rate of 10^{-3} corresponds to an E_s/N_o of about 14.8 dB.

The BPSK link will require a SNR of 7.3 dB to be detected with this level of error performance while the 8-PSK link will require a SNR of 14.8 dB. Now a signal at 95 dBm must be detected and the available power at the antenna is 174 dBm/Hz, which means that in a 1-MHz bandwidth there will be 114 dBm of noise. Thus, the minimum SNR at the antenna is the difference between 95 and 114 dBm, or 19 dB. For the BPSK case, the SNR at the output of the receiver can be 7.3 dB; the SNR can degrade by 11.7 dB. Since $NF = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}$ this means that the *NF* of the radio for the BPSK case can be as high as 11.7 dB. For the 8-PSK case the SNR can only degrade by 19 − 14.8 = 4.2 dB. Thus, in this case the receiver *NF* must be better than 4.2 dB. Note that a receiver with an *NF* of less than 4.2 dB would be quite impressive with typical frequencies and technologies available at the time of writing.

Differential Phase Shift Keying (DPSK)

In the previous section there was one important assumption made about the modulation. It was assumed that it was coherently detected. What this means is that the receiver somehow knew the exact absolute value of the transmitted phase of the signal and used this perfect reference as a comparison to the data that was actually received. Building a circuit that recovers the absolute phase of the received signal is difficult; often a version of PSK is used that does not need any knowledge of the absolute phase of the transmitted signal. This type of modulation is called differential phase shift keying. Differential phase shift keying compares the phase of signal over the current bit period to the phase in the previous bit period. The change (or lack of change) tells the receiver what the value of the bits are that were transmitted. Thus, for instance in DBPSK if there was no change in phase then a zero was transmitted, and if the phase of the carrier changes by 180° then a one was transmitted.

The probability of symbol error for noncoherent detection of DPSK is given by [13]:

$$P_B(M) \approx \frac{2}{\log_2 M} Q\left(\sqrt{2\frac{E_s}{N_o}} \sin\left(\frac{\pi}{\sqrt{2}M}\right)\right) \tag{2.98}$$

This is shown in Figure 2.17.

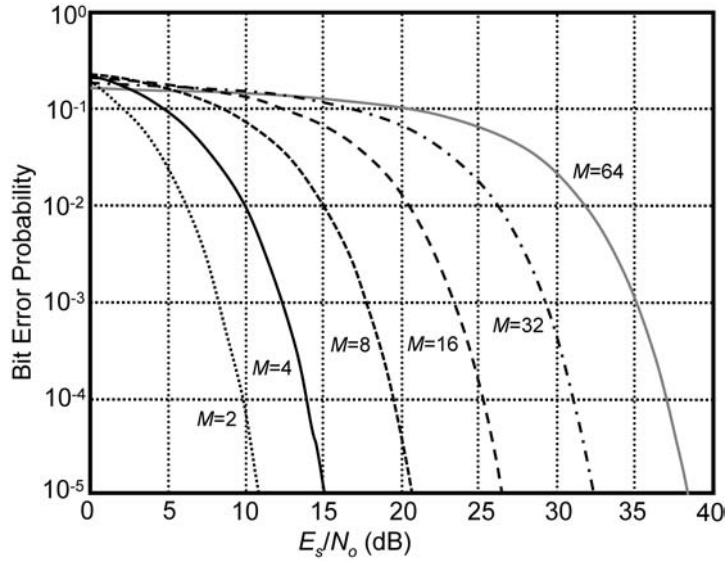

Figure 2.17 Probability of bit error versus E_s/N_o for differential PSK.

2.4.2 Frequency Modulation

In PSK modulation, the phase of the carrier is changed as a means to transmit information. On the other hand, a frequency shift keyed (FSK) modulated signal encodes data by changing the frequency of the carrier signal according to the particular bits that have been transmitted. An FSK signal is discribed by

$$S_{\mathrm{FSK}}(t) = A \times \cos((\omega_{\mathrm{RF}} + \omega_i)t + \phi) \tag{2.99}$$

where A is the amplitude of the carrier signal (a constant), ω_{RF} is the nominal frequency of the carrier, ϕ is an arbitrary phase, and ω_i is the change in carrier frequency that determines what bits have been transmitted. Thus, in this modulation scheme, the RF waveform can be thought of as a set of different frequencies being turned on and off at the bit rate as shown in Figure 2.18.

So, what is to stop us from spacing the different frequencies corresponding to different bits infinitely close together and thus getting an infinite number of

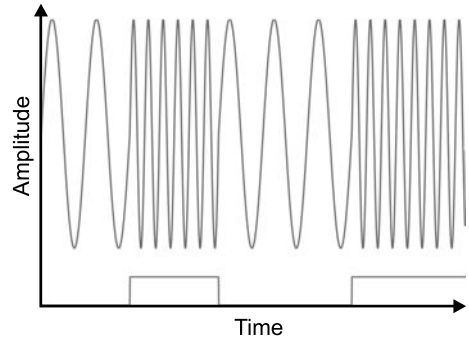

Figure 2.18 Example of a BFSK waveform.

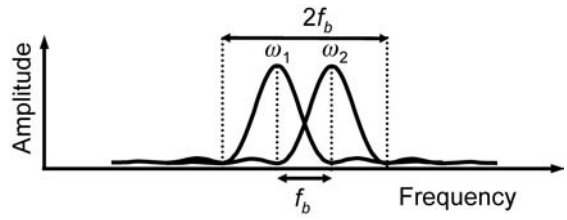

Figure 2.19 Illustration of RF power spectral density for binary FSK.

bits/sec/Hz? Remember that the signals (centered at their carrier frequencies) will have a power spectral density similar to that shown in Figure 2.12. Thus, adjacent frequencies must be spaced so that there is minimal interference between different bits. A way to do this is to align the peak in the frequency response of one bit with the null in the frequency response of an adjacent bit, by spacing the frequencies either at the bit rate or at a multiple of the bit (or symbol) rate as shown in Figure 2.19. Therefore

$$\begin{aligned} \omega_i &= f_B \times i \\ i &= 1,2,3...M \end{aligned} \tag{2.100}$$

where $\log_2 M$ is the number of bits transmitted per frequency. Therefore the maximum frequency deviation of the RF signal is:

$$\omega = f_B \times N \tag{2.101}$$

Thus, the required bandwidth is proportional to the number of bits that are transmitted simultaneously. This means that two bits transmitted simultaneously would take twice the bandwidth that one bit would take. As a result, the spectral efficiency of FSK drops off from around 0.5 bit/sec/Hz for binary FSK and 4-FSK, to lower values as the number of bits increases. Note that even for binary FSK the spectral efficiency is half of that for binary PSK. Therefore, FSK does not offer the same advantage in spectral efficiency offered by PSK. However, since all the frequencies are orthogonal, the presence of more frequencies does not affect the receiver's ability to detect any of them. As a result, higher data rates can be achieved without a reduction in the symbol error rate.

Figure 2.20 shows the probability of a bit error versus E_s/N_o for different numbers of bits per symbols for FSK. Note that the probability of symbol error at a given E_b/N_o actually drops for a higher number of bits per symbol, but also note that this is because a higher bandwidth per symbol is being used. Thus, for higher order FSK, the SNR will need to be higher to get the same E_b/N_o ratio, and as a result, the bandwidth of the radio must be larger. This is different from PSK, where the ratio of bandwidth to the number of bits per symbol is relatively constant, regardless of the number of bits per symbol being transmitted. The probability of symbol error for FSK is given by [13]

$$P_s \left(\frac{E_s}{N_o}, M \right) \leq (M - 1) \times Q\left(\sqrt{\frac{E_s}{N_o}} \right) \tag{2.102}$$

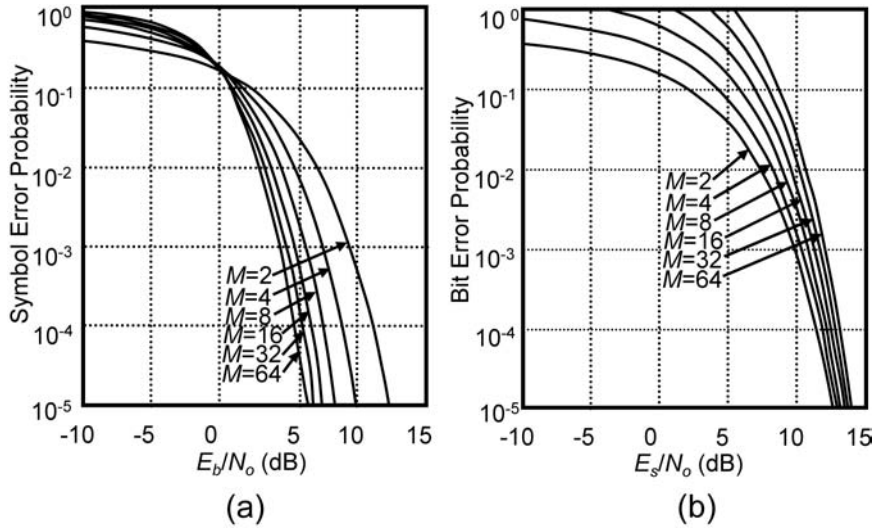

Figure 2.20 Probability of (a) symbol error versus E_b/N_o and (b) bit error versus E_s/N_o for FSK.

Note that with FSK the bit error probability is related to the symbol error probability by:

$$P_B = \frac{M/2}{M-1} P_s \tag{2.103}$$

Therefore the probability of bit error is given by:

$$P_B \left(\frac{E_s}{N_o}, M\right) = \frac{M}{2} Q\left(\sqrt{\frac{E_s}{N_o}}\right) \tag{2.104}$$

2.4.3 Minimum Shift Keying (MSK)

A related modulation that can be thought of as either phase or frequency modulation is called minimum shift keying (MSK). It is a very simple form of modulation that, during each bit period, either advances the phase of the carrier by 90° to indicate a one, or retards the phase by 90° to indicate a zero. Thus, an MSK signal can be represented by:

$$S_{\text{MSK}}(t) = A\cos\left(\omega_{\text{RF}} t + \frac{\pi}{2} d(t) t + \phi\right) \tag{2.105}$$

where $d(t)$ has a value of ±1 to indicate the value of a bit that was transmitted. In order to generate this phase change, the frequency must be instantaneously higher or lower than the carrier frequency over the bit period. Since the phase must change by $\pi/2$ in a period of T_B (the bit rate), the frequency must be

$$f = \frac{\pi/2}{T_B} \frac{1}{2\pi} = \frac{f_B}{4} \tag{2.106}$$

higher or lower than the nominal carrier frequency for that bit period. Note that MSK is very similar to BFSK except that the two frequencies in MSK are spaced at half the separation compared to BFSK. This means that MSK is more spectrally efficient than BFSK. Also, note that MSK is a form of continuous phase modulation. That means that there are no discontinuities in the phase of the transmitted waveform (FSK has no such restriction). This modulation has many properties similar to BPSK and QPSK and will have a bit-error probability that is the same as QPSK as shown in Figure 2.16.

2.4.4 Quadrature Amplitude Modulation (QAM)

Quadrature amplitude modulation (QAM) can be thought of as an extension of QPSK. In QAM, the symbols are distinguished by having both different phases and amplitudes. Thus, instead of four possible phases as in QPSK, a larger number of both phases and amplitudes are used to define which bit has been transmitted. Thus, rather than a constellation of four symbols, in QAM the constellation has 16, 64, 256, or more phase and amplitude locations corresponding to different bits being transmitted. The constellation for 16-QAM is shown in Figure 2.21. In 16-QAM, four bits are transmitted simultaneously for a spectral efficiency of 4 bits/sec/Hz. Similarly, 64-QAM and 256-QAM achieve a spectral efficiency of 6 and 8 bits/sec/Hz respectively.

QAM has the advantage over MPSK in that for a given spectral efficiency, it will often achieve an equivalent bit error rate at a lower E_b/N_o. The probability of bit error for QAM is given by [13]:

$$P_B \left(\frac{E_b}{N_o}, L\right) \approx \frac{2(1 - L^{-1})}{\log_2 L} Q\left(\sqrt{\frac{3\log_2(L)}{L^2 - 1}\frac{2E_b}{N_o}}\right) \tag{2.107}$$

where $L = \sqrt{M}$, which is also the number of amplitude levels in one dimension. So for instance, in the case of 64-QAM, $L = 16$. The bit error rate can also be expressed as:

$$P_B \left(\frac{E_s}{N_o}, L\right) \approx \frac{2(1 - L^{-1})}{\log_2 L} Q\left(\sqrt{\frac{3\log_2(L)}{L^2 - 1}\frac{2E_s}{N_o}\frac{1}{\log_2 L^2}}\right) \tag{2.108}$$

The bit error probability is shown in Figure 2.22.

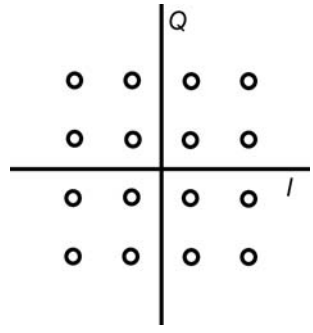

Figure 2.21 Phase plot of a 16-QAM modulated signal.

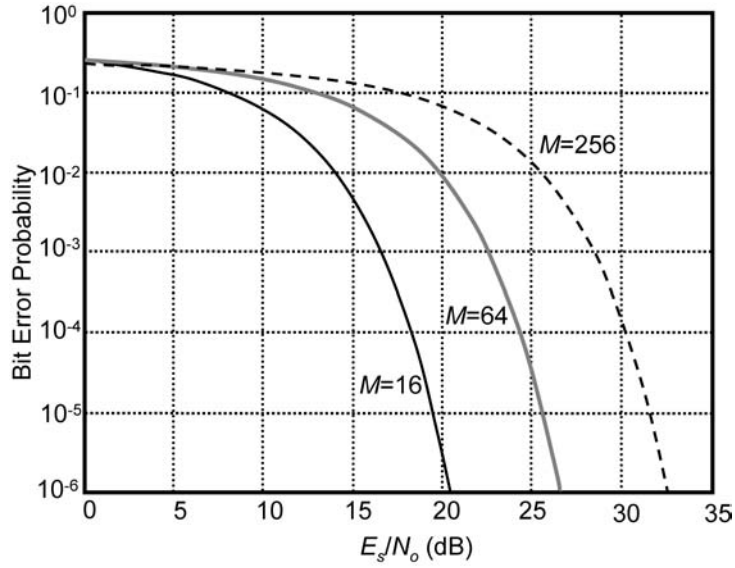

Figure 2.22 Probability of bit error versus E_s/N_o for 16-, 64-, and 256-QAM.

2.4.5 Orthogonal Frequency Division Multiplexing (OFDM)

OFDM is a designed to help improve the link performance in a radio link where there can be frequency selective fading or interference within the bandwidth of the channel. The idea is to replace a single carrier with multiple carriers (called subcarriers), which individually handle data at lower rates. Thus, instead of having one carrier at say 64 Mbps, there can be 64 subcarriers each at 1 Mbps. Usually, the subcarriers are either a form of PSK or QAM modulation. Just as in FSK, the carriers must be spaced at multiples of the data rate they are transmitting. This way nulls of adjacent carriers fall at other carrier center frequencies. This is the reason for "orthogonal" in OFDM, which means loosely that the individual carriers will not interfere with each other. Note that for OFDM, the bit error rate is the same as the single carrier case using the same modulation. Thus for an OFDM system using QPSK the bit error rates previously discussed still apply. The advantage with OFDM is that if interference corrupts one of the subcarriers then only the data on that subcarrier is affected while the others are not. This leads to a lower bit error rate in the presence of frequency selective interference or fading. Often in practical systems certain subcarriers have special functions. They can be used as pilot tones sometimes used to establish the phase of the signal as well as perform many other functions that aid the digital signal processor in determining what data was sent. However, OFDM signals are not constant envelope signals even when using modulations like QPSK, and the signal amplitude can have high peak-to-average ratio.

References

[1] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1984.

[2] Sze, S. M., *Physics of Semiconductor Devices*, 2nd ed., New York: John Wiley & Sons, 1981.

[3] Gray, P. R., et al., *Analysis and Design of Analog Integrated Circuits*, 4th ed., New York: John Wiley & Sons, 2001.

[4] Stremler, F. G., *Introduction to Communication Systems*, 3rd ed., Reading, MA: Addison-Wesley, 1990.

[5] Jordan, E. C., and K. G. Balmain, *Electromagnetic Waves and Radiating Systems*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1968.

[6] Rappaport, T. S., *Wireless Communications*, Upper Saddle River, NJ: Prentice-Hall, 1996.

[7] Proakis, J. G., *Digital Communications*, 3rd ed., New York: McGraw-Hill, 1995.

[8] Gonzalez, G., *Microwave Transistor Amplifiers,* 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 1997.

[9] Wambacq, P., and W. Sansen, *Distortion Analysis of Analog Integrated Circuits*, Boston, MA: Kluwer Academic Publishers, 1998.

[10] Wambacq, P., et al., "High-Frequency Distortion Analysis of Analog Integrated Circuits," *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing*, Vol. 46, No. 3, March 1999, pp. 335–345.

[11] "Some Notes on Composite Second and Third Order Intermodulation Distortions," Matrix Technical Notes MTN-108, December 15, 1998.

[12] "The Relationship of Intercept Points and Composite Distortions," Matrix Technical Notes MTN-109, February 18, 1998.

[13] Sklar, B., *Digital Communications: Fundamentals and Applications*, 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 2001.

CHAPTER 3

System Level Architecture and Design Considerations

3.1 Transmitter and Receiver Architectures and Some Design Considerations

3.1.1 Superheterodyne Transceivers

A block diagram of a typical superheterodyne radio transceiver using air as a medium is shown in Figure 3.1. Modulated signals are transmitted and received at some frequency by an antenna. If the radio is receiving information, then the signals are passed from the antenna to the receiver (Rx) part of the radio. If the radio is transmitting, then signals are passed to the antenna from the transmitter (Tx) part of the radio. Radios either transmit and receive at the same time (called a full duplex transceiver) or alternate between transmitting and receiving (called a half duplex transceiver). In a half duplex transceiver it is possible to put a switch between the antenna and the Rx and Tx paths to provide improved isolation, while in a full duplex transceiver the switch must be omitted and the two input filters have the sole responsibility of isolating the Tx and Rx paths without the aid of a switch.

In the Rx path, the signals are first passed through a filter to remove interference in frequency bands other than the one of interest. The signal is then amplified by a low noise amplifier (LNA) to increase the power in weak signals, while adding as little noise (unwanted random signals) as possible. This amplifier may or may not include some form of adjustable gain or gain steps. The spectrum is then further filtered by an image filter and then downconverted by a mixer to an intermediate frequency (IF). The IF frequency must be chosen with great care, taking into account many factors including interaction of spurs and mixing of LO harmonics.

The mixer (also sometimes called a multiplier) mixes the incoming spectrum of radio-frequency (RF) signals with the output from the RF frequency synthesizer which is an accurate frequency reference generator also called a local oscillator (LO). The LO is tuned so that the frequency of the desired IF signal is always at the same frequency. The LO can be either low-side injected (the LO is at a frequency less than the RF frequency) or high-side injected (the LO is at a frequency greater than the RF frequency). For low-side injection, the IF frequency is given by:

$$f_{IF} = f_{RF} \quad f_{LO} \tag{3.1}$$

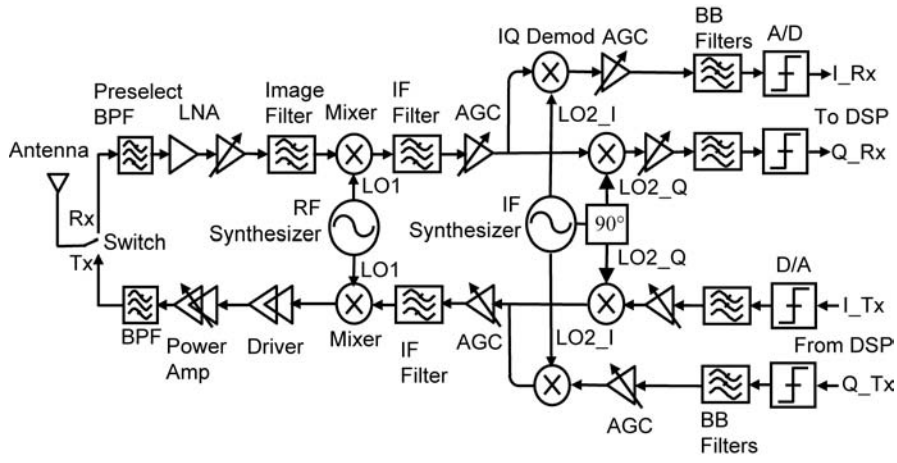

Figure 3.1 A typical half-duplex single IF superheterodyne radio transceiver.

For high-side injection:

$$f_{IF} = f_{LO} \quad f_{RF} \tag{3.2}$$

Since the IF stage is at a fixed frequency, the synthesizer must be made programmable so that it can be tuned to whatever input frequency is desired at a given time. An input signal at an equal distance from the LO on the other side from the desired RF signal is called the image signal. It is called the image because a signal at this frequency after mixing will be at the same IF as the desired signal. Therefore, the image signal cannot be removed by filtering after mixing has taken place. Thus, an important job of the RF filters is to remove any image signals before such mixing action takes place. This is illustrated in Figure 3.2.

After mixing to an IF, additional filtering is usually performed. At the IF, unwanted channels can be filtered out leaving the channel of interest now centered at

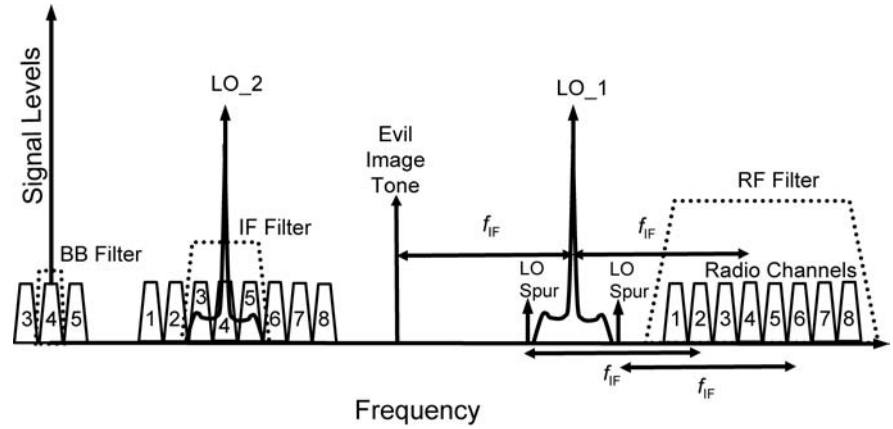

Figure 3.2 Figure showing radio receiver frequency plan. As shown, the radio is tuned to receive channel 4 and has a low-side injected LO.

the IF, and possibly some subset of the original adjacent channels depending on the quality of the filter used. Usually, automatic gain control (AGC) amplifiers are also included at the IF. They adjust the gain of the radio so that its output amplitude is always constant. Once through the AGC, the signals are downconverted a second time to baseband (the signals are now centered around DC or zero frequency). This second downconversion requires a second frequency synthesizer that produces both 0° and 90° output signals at the IF frequency. Two mixers downconvert the signals into an in-phase (I) and quadrature phase (Q) paths. By using two separate phases, both amplitude and phase information can be recovered, and as a result, the incoming phase of the RF signal does not need to be synchronized to the phase of the LO tone. The I and Q signals are then passed through baseband filters which removes the rest of the unwanted channels. Finally, the signal is passed through an analog-to-digital converter and into the back end of the radio. There may be additional, possibly programmable, gain stages in the baseband. Further signal processing is performed in the digital signal processing (DSP) circuitry in the back end of the radio.

The transmitter works much the same way except in reverse. The DSP circuitry followed by the digital-to-analog converters (DA) produces signals in quadrature. These signals are then filtered and upconverted to an IF frequency. The Tx will usually have some AGC function, which may be either in the baseband or the IF stage. The IF signal is upconverted to the RF frequency by the mixer. If the LO is low-side injected, the mixer is used to generate sum rather than difference products. Thus, for low-side injection, the RF frequency is given by:

$$f_{RF} = f_{LO} + f_{IF} \tag{3.3}$$

If the LO is high-side injected, the frequency of the RF signal is given by:

$$f_{RF} = f_{LO} \quad f_{IF} \tag{3.4}$$

Once upconverted to RF, the signal is passed through a power amplifier to increase the power of the signals and is then radiated by the antenna into the air. In the RF section the PA itself may have a power control function or additional AGC. If the power level is constant, it must be high enough so that the signal can be detected at the maximum distance dictated by the system specifications. The mixer will produce signals on each side of the LO. The RF BPF is needed to filter the signal so that only the sideband at the desired RF transmit frequency is passed to the antenna. The BPF can also be used to remove LO feedthrough from the mixer.

3.1.2 Direct Conversion Transceivers

A direct downconversion radio architecture is shown in Figure 3.3. In this architecture, the IF stage is omitted and the signals are converted directly to DC as shown in Figure 3.4. For this reason the architecture is sometimes called a zero-IF radio. The direct conversion transceiver saves the area and power associated with a second synthesizer. No image filter is required and the LO frequency selection becomes trivial. However, generating I and Q signals from a synthesizer at higher

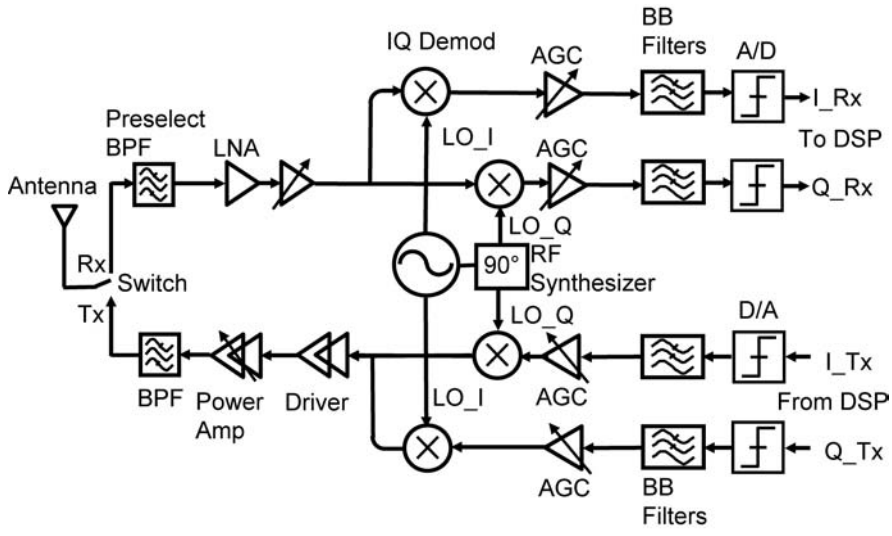

Figure 3.3 A typical half-duplex direct downconversion radio.

frequencies is much more difficult than doing so at the IF. Since the LO signal is now at the same frequency as the incoming RF signal, LO energy can couple into the RF path and cause problems. Without an IF stage, more gain and gain control must be done at baseband making the amplitude and phase matching of both the I and Q paths more difficult than in the case of the superheterodyne radio. The transmit path is also simpler than the superheterodyne case. Here the signal is converted directly to RF. This may mean that part or all of the AGC function will need to be completed at RF to avoid matching problems. However, in this case the RF filter requirements on the transmitter are greatly reduced since direct conversion means there is no unwanted sideband to be filtered out or LO feedthrough to worry about.

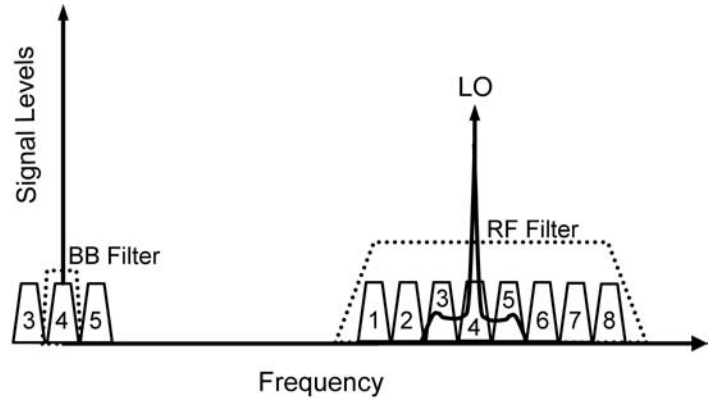

Figure 3.4 Figure showing radio receiver frequency plan. As shown, the radio is tuned to receive channel 4.

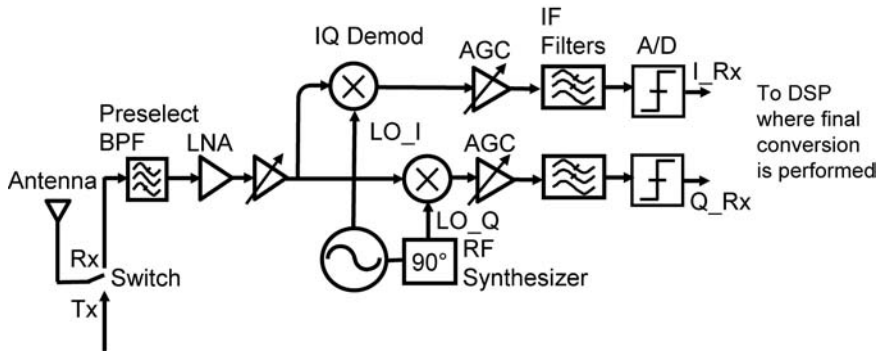

Figure 3.5 A low IF receiver.

3.1.3 Low IF Transceiver and Other Alternative Transceiver Architectures

One alternative architecture is called a low IF transceiver. The low IF architecture sees most of its advantage on the receive side. While this architecture is conceptually the same as a superheterdyne radio, the low IF is chosen so that it is practical to digitize the signal at the IF and thus perform some of the traditionally analog signal processing in the back end of the radio. The main problem with a low IF receiver is that the image frequency will now be very close to the desired sideband. However, if the signal is then digitized, it is easier to employ a high Q digital filter to remove the image or alternatively to make use of an image reject mixer. Image reject mixers will be studied in Chapter 8. An example of a low IF receiver is shown in Figure 3.5.

Another architecture that is a compromise between the superheterodyne and the direct-conversion transceiver is called a walking IF architecture shown in Figure 3.6. This architecture derives the IF LO by dividing the RF LO by some fixed number. As a result, the IF frequency is not fixed but "walks" in step with a fraction of the frequency of the RF LO. This transceiver with walking IF still has many of the advantages of the superheterodyne radio (although it is not possible to filter as

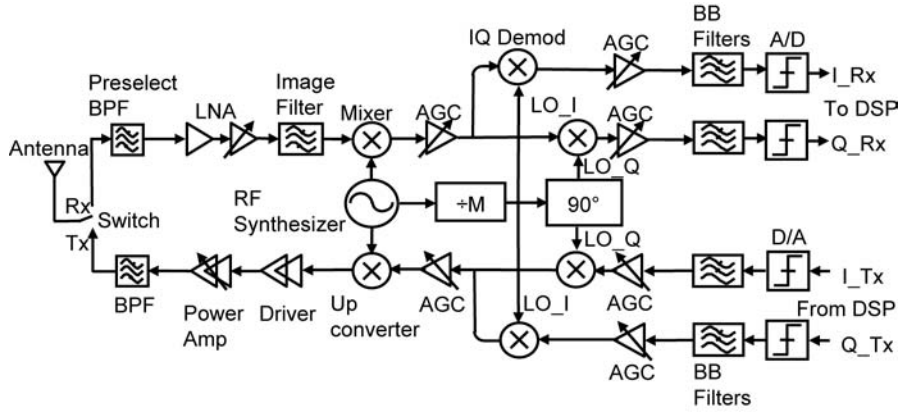

Figure 3.6 A walking IF radio architecture.

well at IF), but also removes the need for the extra synthesizer, potentially reducing layout area and power dissipation.

3.2 System Level Considerations

The next subsections will discuss some of the most important design considerations when specifying the requirements of all the components in the system. For additional information the reader should consider [1–7].

3.2.1 The Noise Figure of Components in Series

For components in series, as shown below in Figure 3.7, one can calculate the total output noise $N_{o(\text{total})}$ and output noise due to the source $N_{o(\text{source})}$ to determine the noise figure.

The output signal S_o is given by:

$$S_o = S_i \times G_1 \times G_2 \times G_3 \tag{3.5}$$

The input noise is:

$$N_{i(\text{source})} = kT \tag{3.6}$$

The total output noise is:

$$N_{o(\text{total})} = N_{i(\text{source})}G_1G_2G_3 + N_{o1(\text{added})}G_2G_3 + N_{o2(\text{added})}G_3 + N_{o3(\text{added})} \tag{3.7}$$

The output noise due to the source is:

$$N_{o(\text{source})} = N_{i(\text{source})}G_1G_2G_3 \tag{3.8}$$

Finally, the noise factor can be determined as:

$$\begin{aligned} F = \frac{N_{o(\text{total})}}{N_{o(\text{source})}} &= 1 + \frac{N_{o1(\text{added})}}{N_{i(\text{source})}G_1} + \frac{N_{o2(\text{added})}}{N_{i(\text{source})}G_1G_2} + \frac{N_{o3(\text{added})}}{N_{i(\text{source})}G_1G_2G_3} \\ &= F_1 + \frac{F_2 \quad 1}{G_1} + \frac{F_3 \quad 1}{G_1G_2} \end{aligned} \tag{3.9}$$

The above formula shows how the presence of gain preceding a stage causes the effective noise figure to be reduced compared to the measured noise figure of

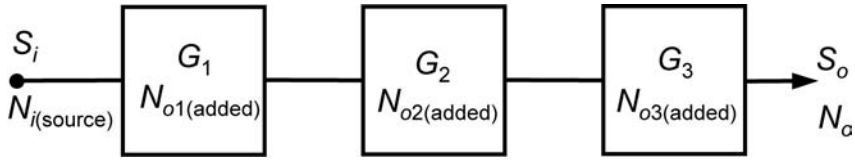

Figure 3.7 Noise figure in cascaded circuits with gain and noise added shown in each.

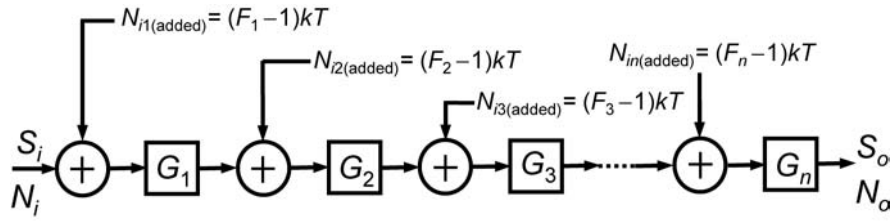

Figure 3.8 Equivalent noise model of a circuit.

a stage by itself. For this reason, we typically design systems with a low noise amplifier at the front of the system. We note that the noise figure of each block is typically determined for the case in which a standard input source (e.g., 50) is connected. The above formula can also be used to derive an equivalent model of each block as shown in Figure 3.8. If the input noise when measuring noise figure is:

$$N_{i(\text{source})} = kT \tag{3.10}$$

and noting from manipulation of (2.14) that:

$$N_{o(\text{added})} = (F \quad 1)N_{o(\text{source})} \tag{3.11}$$

Now dividing both sides of (3.11) by G:

$$N_{i(\text{added})} = (F \quad 1)\frac{N_{o(\text{source})}}{G} = (F \quad 1)N_{i(\text{source})} = (F \quad 1)kT \tag{3.12}$$

Thus, the input referred noise model for cascaded stages as shown in Figure 3.8 can be derived.

Example 3.1: Cascaded Noise Figure and Sensitivity Calculation

Find the effective noise figure and noise floor of the system shown in Figure 3.9. The system consists of a filter with 3-dB loss, followed by a switch with 1 dB loss, an LNA, and a mixer. Assume the system needs an SNR of 7 dB for a bit error rate of 10 ^{3}. Also, assume that the system bandwidth is 200 kHz.

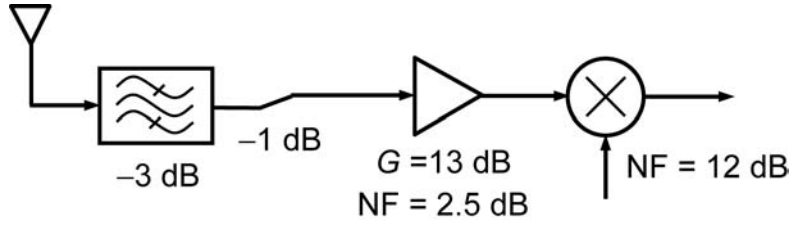

Figure 3.9 System for performance calculation.

Solution:
Since the bandwidth of the system has been given as 200 kHz, the noise floor of the system can be determined:

$$\text{Noise Floor} = \quad 174\ \text{dBm} + 10\log_{10}(200{,}000) = \quad 121\ \text{dBm}$$

We make use of the cascaded noise figure equation and determine that the overall system noise figure is given by:

$$\text{NF}_{\text{TOTAL}} = 3\ \text{dB} + 1\ \text{dB} + 10\log_{10}\ \ 1.78 + \frac{15.84 \quad 1}{20} \qquad 8\ \text{dB}$$

Note that the LNA noise figure of 2.5 dB corresponds to a noise factor of 1.78 and the gain of 13 dB corresponds to a power gain of 20. Furthermore, the noise figure of 12 dB corresponds to a noise factor of 15.84.

Note that if the mixer also has gain, then possibly the noise due to the IF stage may be ignored. In a real system this would have to be checked, but here we will ignore noise in the IF stage.

Since it was stated that the system requires an S/N ratio of 7 dB, the sensitivity of the system can now be determined:

$$\text{Sensitivity} = \quad 121\ \text{dBm} + 7\ \text{dB} + 8\ \text{dB} = \quad 106\ \text{dBm}$$

Thus, the smallest allowable input signal is 106 dBm. If this is not adequate for a given application, then a number of things can be done to improve this:

A smaller bandwidth could be used. This is usually fixed by IF requirements.
The loss in the preselect filter or switch could be reduced. For example, the LNA could be placed in front of one or both of these components.
The noise figure of the LNA could be improved.
The LNA gain could be increased reducing the effect of the mixer on the system NF.
A lower NF in the mixer would also improve the system NF.
If a lower SNR for the required BER could be tolerated, then this would also help.

In a modern completely integrated radio, often only the interface with off-chip circuitry has a standard impedance and the interfaces between blocks on chip tend to be undefined. In this case, defining a noise figure for each sub-block is less convenient. On chip most blocks may be better characterized by a loaded voltage gain G_{vi} and an added noise voltage. In this case, the overall input referred noise for the system can be given as:

$$v_{ni(added)}{}^2 = v_{n1}{}^2 + \frac{v_{n2}}{G_{v1}{}^2} + \frac{v_{n3}}{G_{v1}{}^2 G_{v2}{}^2} + \ldots \tag{3.13}$$

where v_{n1}, v_{n2}, and v_{n3} are the input referred noise voltages of the three stages. If we assume that the input of the first block is matched to the source resistance then the noise figure is given by [8]:

$$F = \frac{N_{i(\text{total})}}{N_{i(\text{source})}} = \frac{\dfrac{\upsilon_{ni(added)}{}^2 + \upsilon_{ni(source)}{}^2}{R_s}}{\dfrac{\upsilon_{ni(source)}{}^2}{R_s}} = 1 + \frac{\upsilon_{n1}{}^2 + \dfrac{\upsilon_{n2}{}^2}{G_{v1}{}^2} + \dfrac{\upsilon_{n3}{}^2}{G_{v1}{}^2 G_{v2}{}^2} + \ldots}{\upsilon_{ni(source)}{}^2}$$

$$= 1 + \frac{\upsilon_{n1}{}^2 + \dfrac{\upsilon_{n2}{}^2}{G_{v1}{}^2} + \dfrac{\upsilon_{n3}{}^2}{G_{v1}{}^2 G_{v2}{}^2} + \ldots}{kTR_s} \tag{3.14}$$

$$= 1 + \frac{\upsilon_{n1}{}^2}{kTR_s} + \frac{\upsilon_{n2}{}^2}{kTR_s G_{v1}{}^2} + \frac{\upsilon_{n3}{}^2}{kTR_s G_{v1}{}^2 G_{v2}{}^2} + \ldots$$

This formula would provide the system designer with the opportunity to specify voltage gain and input referred noise voltage rather than noise figure, which makes more sense.

Example 3.2: On-Chip NF Confusion

A system level designer unfamiliar with IC design has set the system level requirements for a radio. The system designer has made the mistake of assuming that all blocks are matched to 50 . The RF front end has a gain of 10 dB and an NF of 4 dB. The first block in the IF stage is a variable gain amplifier. The system level designer specifies this block to have an NF of 10 dB so that the overall NF of the LNA, mixer, and VGA will be:

$$F = F_{LNA+Mix} + \frac{F_{VGA} \quad 1}{G_{LNA+Mix}} = 2.51 + \frac{10 \quad 1}{10} = 3.41 = 5.33 \text{ dB}$$

This is below the target for this design of 6.5 dB and makes sure that the VGA is not a dominant concern. Once design is underway, the mixer designer adds a buffer with an output impedance of 5 to drive the VGA, which has an input impedance of 5 k . The VGA designer simulates his block using a 50 source and finds he has an NF of 15 dB. Assuming the VGA input can be modeled with the 5-k input resistance parallel to input noise current, does he meet the intended noise requirement for the block?

Solution:
Appling the same formula as the system level designer initially used, the noise figure of the system with a VGA with a 15-dB noise figure would be:

$$F = F_{LNA+Mix} + \frac{F_{VGA} \quad 1}{G_{LNA+Mix}} = 2.51 + \frac{31.6 \quad 1}{10} = 5.57 = 7.46 \text{ dB}$$

This is about 1 dB over the spec, so it is clearly a concern for the system designer. The noise voltage at the VGA input with a 50 input is:

$$N_{VGA} = \frac{v_{nVGA}^2}{R_s} = (F \quad 1)kT \qquad v_{nVGA} = \sqrt{(F \quad 1)kTR_s} = \sqrt{\frac{30.6}{4} \quad 4kTR_s}$$

$$= 2.76 \quad 0.9 \text{ nV} = 2.49 \text{ nV}/\sqrt{\text{Hz}}$$

If the same noise current is now fed into $5 \quad \| 5 \text{ k} \quad$ instead of $50 \quad \| 5 \text{ k} \quad$, the noise voltage will be approximately 1/10 since in either case the 5 k resistor can be ignored. So using $V_{nVGA} = 0.249$ nV results in:

$$F = F_{RF} + \frac{{v_{nVGA}}^2}{kTR_sG_{vRF}^2} = 2.51 + \frac{4 \quad (0.249 \text{ nV})^2}{(0.9 \text{ nV})^2 \quad 10} = 2.51 + 0.03 = 2.54 = 4.05 \text{ dB}$$

So the VGA has nearly no effect on noise, but a real VGA would have both current and voltage noise, leading to somewhat higher noise. Also, noise due to an extra buffer is low due to the G_{RF} preceding it. Note that G_{RF} is power gain and G_{vRF} is voltage gain. So in this case we see that even though the block looks like it fails the spec, it actually only adds a small fraction of the noise allowed by the original specification. However, because 15 dB is more than the 10 dB specified, this now results in pressure on the VGA designer by the system designer possibly resulting in wrongful termination and a series of lawsuits. If only the systems designer had used (3.14), life would have been so much better for all involved!

3.2.2 The Linearity of Components in Series

Linearity can also be computed for components in series. Starting with two amplifiers in series that have unique nonlinear transfer functions of:

$$\begin{aligned} v_{o1} &= k_{a1}v_i + k_{a2}{v_i}^2 + k_{a3}{v_i}^3 \\ v_{o2} &= k_{b1}v_{o1} + k_{b2}{v_{o1}}^2 + k_{b3}{v_{o1}}^3 \end{aligned} \tag{3.15}$$

Each block will have an IIP3 of:

$$\begin{aligned} v_{\text{IIP3\_1}} &= 2\sqrt{\frac{k_{a1}}{3 \mid k_{a3} \mid}} \\ v_{\text{IIP3\_2}} &= 2\sqrt{\frac{k_{b1}}{3 \mid k_{b3} \mid}} \end{aligned} \tag{3.16}$$

From (3.15) v_{o2} can be expanded to give an overall transfer function of:

$$v_{o2} = k_{a1}k_{b1}v_i + (k_{a1}k_{b1} + {k_{a1}}^2k_{b2})v_i^2 + (k_{b1}k_{a3} + 2k_{b2}k_{a1}k_{a2} + k_{b3}{k_{a1}}^3)v_i^3 + \ldots \tag{3.17}$$

Note that this equation is truncated after the third-order terms and it is assumed that higher order terms are less important.

Now applying the definition of IIP3 to the overall transfer function yields:

$$v_{\mathrm{IIP3}} = 2\sqrt{\frac{k_{a1}k_{b1}}{3\,|\,k_{b1}k_{a3} + 2k_{b2}k_{a1}k_{a2} + k_{b3}{k_{a1}}^3\,|}} \tag{3.18}$$

Now if both sides are squared and inverted this becomes:

$$\frac{1}{{v_{\mathrm{IIP3}}}^2} = \frac{3}{4}\times\frac{|k_{b1}k_{a3}| + |2k_{b2}k_{a1}k_{a2}| + \left|k_{b3}{k_{a1}}^3\right|}{k_{a1}k_{b1}} = \frac{3}{4}\times\frac{|k_{b1}k_{a3}|}{k_{a1}k_{b1}} + \frac{|2k_{b2}k_{a1}k_{a2}|}{k_{a1}k_{b1}} + \frac{\left|k_{b3}{k_{a1}}^3\right|}{k_{a1}k_{b1}} \div \div \tag{3.19}$$

If it is assumed that the second-order terms k_{a2} and k_{b2} are small compared to the first-order terms this expression can be simplified:

$$\frac{1}{{v_{\mathrm{IIP3}}}^2} = \frac{1}{{v_{\mathrm{IIP3\_1}}}^2} + \frac{3k_{b2}k_{a2}}{2k_{b1}} + \frac{{k_{a1}}^2}{{v_{\mathrm{IIP3\_2}}}^2} \quad \frac{1}{{v_{\mathrm{IIP3\_1}}}^2} + \frac{{k_{a1}}^2}{{v_{\mathrm{IIP3\_2}}}^2} \tag{3.20}$$

In general, the IIP3 of cascaded stages with voltage gain A_v is given by:

$$\frac{1}{{v_{\mathrm{IIP3}}}^2} = \frac{1}{{v_{\mathrm{IIP3\_1}}}^2} + \frac{{A_{v1}}^2}{{v_{\mathrm{IIP3\_2}}}^2} + \frac{{A_{v1}}^2{A_{v2}}^2}{{v_{\mathrm{IIP3\_3}}}^2} + \ldots \tag{3.21}$$

Since (3.21) uses the square of the voltage, assuming all blocks are matched to
50 it can be rewritten in terms of power:

$$\frac{1}{IIP3} = \frac{1}{IIP3\_1} + \frac{G_1}{IIP3\_2} + \frac{G_1G_2}{IIP3\_3} + \ldots \tag{3.22}$$

where G_1 and G_2 are the power gains of the respective blocks.

Similarly the 1 dB compression point of a cascaded system would be:

$$\frac{1}{{v_{\mathrm{1dB}}}^2} = \frac{1}{{v_{\mathrm{1dB\_1}}}^2} + \frac{{A_{v1}}^2}{{v_{\mathrm{1dB\_2}}}^2} + \frac{{A_{v1}}^2{A_{v2}}^2}{{v_{\mathrm{1dB\_3}}}^2} + \ldots \tag{3.23}$$

IIP2 for series blocks has a slightly different form. Using the two element system discussed above gives:

$$v_{\mathrm{IIP2}} = \frac{k_{a1}k_{b1}}{k_{a2}k_{b1} + {k_{a1}}^2k_{b2}} \tag{3.24}$$

Inverting this equation gives:

$$\frac{1}{v_{\mathrm{IIP2}}} = \frac{k_{a2}k_{b1} + {k_{a1}}^2k_{b2}}{k_{a1}k_{b1}} = \frac{k_{a2}}{k_{a1}} + \frac{k_{a1}k_{b2}}{k_{b1}} = \frac{1}{v_{\mathrm{IP2\_1}}} + \frac{k_{a1}}{v_{\mathrm{IP2\_2}}} \tag{3.25}$$

In general, the IIP2 of cascaded stages will be given by:

$$\frac{1}{v_{\text{IIP2}}} = \frac{1}{v_{\text{IIP2\_1}}} + \frac{A_{v1}}{v_{\text{IIP2\_2}}} + \frac{A_{v1}A_{v2}}{v_{\text{IIP2\_3}}} + \ldots \tag{3.26}$$

3.2.3 Dynamic Range

So far, we have discussed noise and linearity in circuits. Noise determines how small a signal a receiver can handle, while linearity determines how large a signal a receiver can handle. If operation up to the 1-dB compression point is allowed (for about 10% distortion, or IM3 is about 20 dB with respect to the desired output), then the dynamic range is from the minimum detectable signal to this point. This is illustrated in Figure 3.10. In this figure, intermodulation components are above the minimum detectable signal for $P_{\text{in}} > -32$ dBm, for which $P_{\text{out}} = -23$ dBm. Thus, for any P_{out} between the minimum detectable signal of -105 dBm and -23 dBm, no intermodulation components can be seen, the so called spurious free dynamic range is 82 dB.

Example 3.3: Determining Dynamic Range

In Example 3.1 we determined the sensitivity of a receiver system. Figure 3.11 shows this receiver again with the linearity of the mixer and LNA specified. Determine the dynamic range of this receiver.

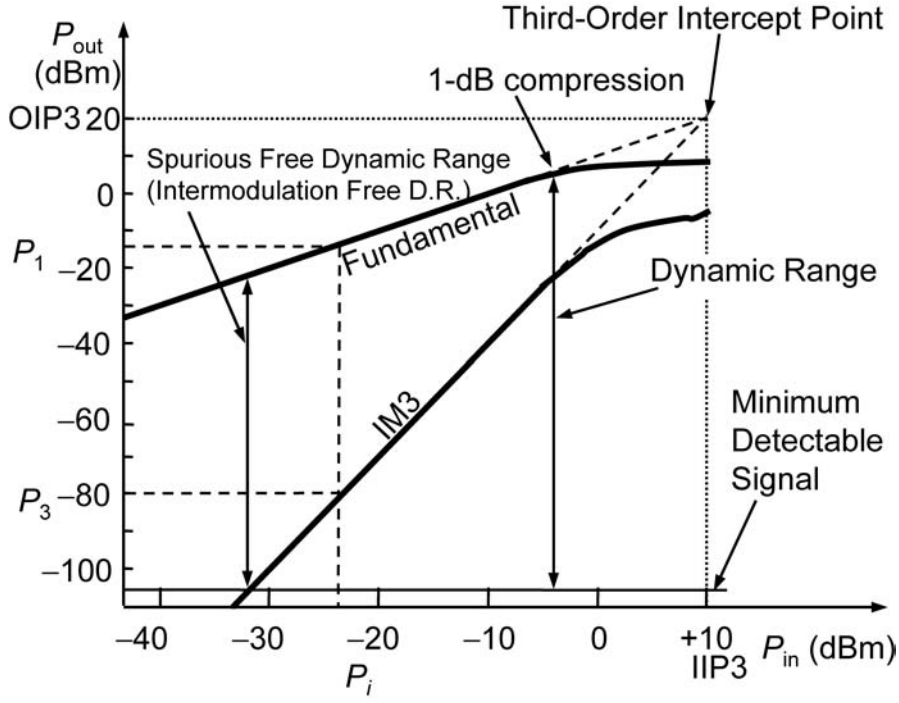

Figure 3.10 Plot of input output power of fundamental and IM3 versus input power showing dynamic range.

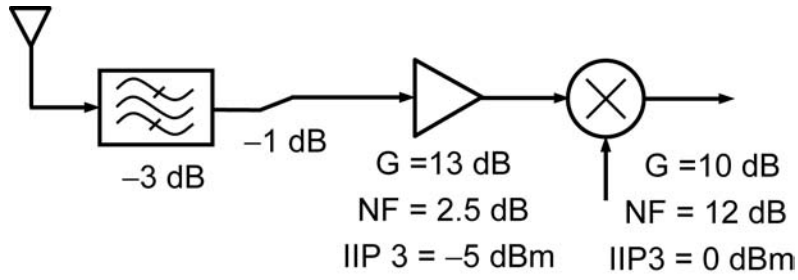

Figure 3.11 System example circuit.

Solution:
The overall receiver has a gain of 19 dB. The minimum detectable signal from Example 3.1 is 106 dBm or 87 dBm at the output. The IIP3 of the LNA and mixer combination from (3.22) is

$$\frac{1}{\text{IIP3}} = \frac{1}{316.2\ \mu\text{W}} + \frac{20}{1\ \text{mW}} = 2.316 \times 10^3$$

$$\text{IIP3} = 43.2\ \mu\text{W} \qquad 13.6\ \text{dBm}$$

This referred to the input is –13.6 + 4 = 9.6 dBm. The IIP3 of the mixer by itself (with a perfect LNA) referred to the input would be 0 – 13 + 4 = 9 dBm. This is close to the IIP3 with the real LNA, therefore, the mixer dominates the IIP3 for the receiver as expected. The 1 dB compression point will be 9.6 dB lower than the 9.6 dBm calculated with the real LNA, or –19.2 dBm. Thus, the dynamic range of the system will be –19.2 + 106 = 86.8 dB.

Example 3.4: Effect of Bandwidth on Dynamic Range

The data transfer rate of the previous receiver can be greatly improved if we use a bandwidth of 80 MHz rather than 200 kHz. What does this do to the dynamic range of the receiver?

Solution:
This system is the same as the last one except that now the bandwidth is 80 MHz. Thus, the noise floor is now:

$$\text{Noise Floor} = 174\ \text{dBm} + 10\log_{10}(80 \cdot 10^6) = 95\ \text{dBm}$$

Assuming that the same signal to noise ratio is required:

$$\text{Sensitivity} = 95\ \text{dBm} + 7\ \text{dB} + 8\ \text{dB} = 80\ \text{dBm}$$

Thus, the dynamic range is now –19.2 + 80 = 60.8 dB. In order to get this back to the value in the previous system we would need to increase the linearity of the receiver by 26 dB! As we will see in future chapters, this would be no easy task.

3.2.4 Image Signals and Image Reject Filtering

At the RF frequency, there are filters to remove out of band signals that may be picked up by the antenna. Any filter in the RF section of the radio must be wide enough to pass the entire receive band, and therefore can do little about any in-band interferers. In a superheterodyne receiver, the filters in the RF section also have the added task of removing the image frequency and are thus sometimes called image reject filters. A superheterodyne receiver takes the desired RF input signal and mixes it with some reference signal to extract the difference frequency. The LO reference is mixed with the input to produce a signal at the difference frequency of the LO and RF as shown in Figure 3.2. As mentioned earlier, a signal on the other side of the LO at the same distance from the LO will also mix down "on top" of the desired frequency. Thus, before mixing can take place, this unwanted image frequency must be removed.

Thus, another important specification in a receiver is how much image rejection it has. Image rejection is defined as the ratio of the gain G_{sig} of the desired signal to the gain of the image signal G_{im}.

$$\mathrm{IR} = 10\log\left(\frac{G_{sig}}{G_{im}}\right) \tag{3.27}$$

In general a receiver must have an image rejection large enough so that in the case of the largest possible image signal and the weakest receive channel power, the ratio of the channel power to the image power, once downconverted, is still larger than the minimum required SNR for the radio.

The amount of filtering can be calculated by knowing the undesired frequency with respect to the filter center frequency, the filter bandwidth, and filter order. The following equation can be used for this calculation:

$$A_{\mathrm{dB}} = \frac{n}{2}\ 20\log\left(\frac{f_{ud}\ \ f_c}{f_{be}\ \ f_c}\right) = \frac{n}{2}\ 20\log\ 2\left(\frac{f}{f_{\mathrm{BW}}}\right) \tag{3.28}$$

where A_{dB} is the attenuation in decibels, n is the filter order, and thus $n/2$ is the effective order on each edge, f_{ud} is the frequency of the undesired signal, f_c is the filter center frequency, and f_{be} is the filter band edge.

Example 3.5: Image Reject Filtering

A system has an RF band from 902–928 MHz, a 200-kHz channel bandwidth, and channel spacing. The first IF is at 70 MHz. Determine the required order of the 26-MHz filter to get a worst-case image rejection of better than 50 dB. If the received image signal power is 40 dBm, the minimum input signal power is 75 dBm, and the required SNR for this modulation is 9.5 dB will this be enough image rejection?

Solution:

The frequency spectrum is shown in Figure 3.12. At RF, the local oscillator frequency f_{LO} is tuned to be 70 MHz above the desired RF signal so that the desired

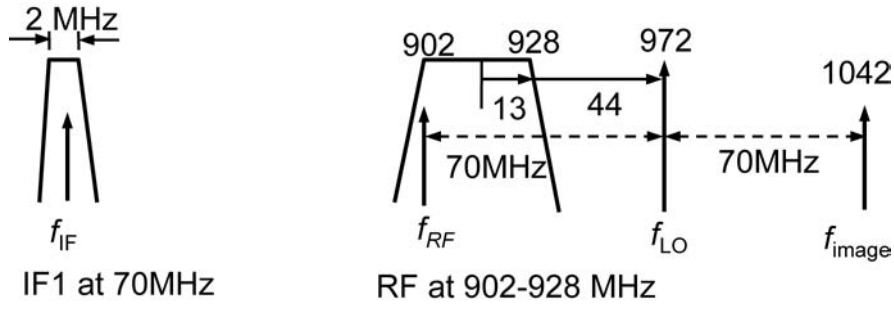

Figure 3.12 Signal spectrum for filter example.

signal will be mixed down to IF1 at 70 MHz. Thus, f_{LO} is adjustable between 972 MHz and 998 MHz to allow signals between 902 MHz and 928 MHz to be received. The image occurs 70 MHz above f_{LO}. The worst case will be when the image frequency is closest to the filter frequency. This occurs when the input is at 902 MHz, the LO is at 972 MHz, and the image is 1,042 MHz. The required filter order n can be calculated by solving (3.28) using f_{BW} = 26 MHz and f = 70 + 44 + 13 = 127 MHz as follows:

$$n = \frac{2 \quad A_{dB}}{20 \quad \log(2 \quad f/f_{BW})} = 5.05$$

Since the order must be an even number, a sixth order filter is used and total attenuation is calculated to be 59.4 dB.

Now with the original image rejection specification, the 40-dBm image signal will be attenuated by 50 dB so after the filter, this signal will be 90 dBm. This will be mixed on top of the desired signal, which has a minimum power level of 75 dBm, giving a signal to distortion ratio of 75 dBm – (90 dBm) = 15 dB. This is higher than the required SNR of 9.5 dB so this should be enough image rejection.

3.2.5 Blockers and Blocker Filtering

The receiver must be able to maintain operation and detect the desired signal in the presence of other signals, often referred to as blockers. These other signals could be of large amplitude and could be close by in frequency; for example, they could be signals in other channels of the same radio band. This is an example of the so-called near-far problem that occurs when the desired signal is far away and one or more interfering signals are close by, which are larger than the wanted signal. A large blocker must not saturate the radio and therefore the 1-dB compression point of the radio must be higher than the blocker power level to avoid saturating the radio.

The intermodulation products of blockers can also be a very big problem. Consider the case where a desired channel is detected at its minimum power level. Two close by channels are also received at their maximum receive power. If these signals are at frequencies such that their IM3 products fall on top of the desired signal, they

will act to reduce the signal-to-noise ratio of the desired channel and could cause an increase in bit error rate. Therefore the circuits in the radio must have a sufficiently high linearity so that this does not happen. Once the received band is down converted to an IF or baseband frequency, filters may be added with a pass band narrower than the whole radio band. As a result, strong adjacent channel signals are filtered; this will reduce the linearity requirements of blocks after the filter.

Example 3.6: How Blockers Are Used to Determine Linearity

Consider typical blocker specifications for a receiver shown in Figure 3.13(a). The input signal is at 102 dBm and the required signal-to-noise ratio, with some safety margin, is 11 dB. Calculate the required input linearity of the receiver. If the receiver front end has 20 dB of gain, is followed by an IF filter with 30 dB of attenuation at offsets between 500 kHz and 2.5 MHz, and has an attenuation of 50 dB at offsets greater than 2.5 MHz as shown in Figure 3.13(b), what is the required linearity of the first circuit in the IF stage?

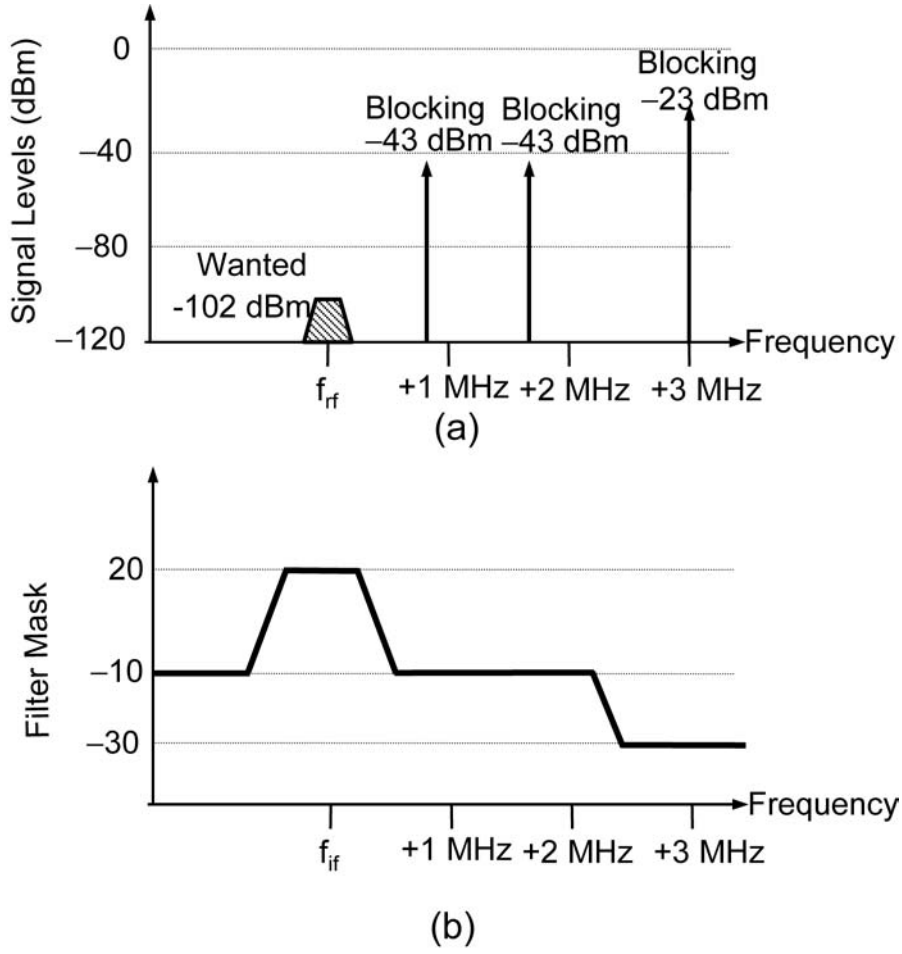

Figure 3.13 (a) Minimum detectable signal and blocker levels at the RF input of the radio, and (b) IF filter mask to help reduce the effect of the blockers.

Solution:
With nonlinearity, third-order intermodulation between the pair of blockers will cause interference directly on top of the signal. The level of this disturbance must be low enough so that the signal can still be detected. The other potential problem is that the large blocker at –23 dBm can cause the amplifier to saturate, rendering the amplifier helpless to respond to the desired signal, which is much smaller. In other words, the receiver has been blocked.

The blocker inputs at –43 dBm will result in third-order intermodulation components, which must be less than –113 dBm (referred to the input) so there is still 11 dB of SNR at the input. Thus, the third-order components (at –113 dBm) are 70 dB below the fundamental components (at –43 dBm). Using (2.59) with P_i at –43 dBm and $[P_1 - P_3]$ = 70 dB results in IIP3 of about 8 dBm, and a 1-dB compression point of about –18 dBm at the input. Thus, the single input blocker at –23 dBm, is still 5 dB away from the 1-dB compression point. This sounds safe—however, there will now be gain through the LNA and the mixer. The blocker will not be filtered until after the mixer, so one must be careful not to saturate any of the components along this path.

Now after the signal passes through the front end and is downconverted to the IF and passed through the IF filter the spectrum will be as shown in Figure 3.14. In this case, the signal experiences a 20 dB gain, while the two closest blockers experience a net gain of 10 dB and the third blocker experiences a net gain of 30 dB. If no filtering were applied to the system then the IIP3 of the first IF block would need to be roughly 8 dBm + 20 dB = 12 dBm. With filtering, the IM3 products from the two closest blockers must be lower than 93 dBm. Using (2.59), with P_i at –53 dBm and $[P_1 - P_3]$ = 40 dB, results in IIP3 of about 33 dBm and a 1-dB compression point of about –43 dBm for the IF block. Thus it is easy to see the dramatic reduction in required linearity with the use of filters. Note that the 1-dB compression point is still 10 dB above the level of any of the blocking tones.

3.2.6 The Effect of Phase Noise on SNR in a Receiver

The blocking signals can cause problems in a receiver through another mechanism known as reciprocal mixing. For a blocker at an offset of f from the desired signal,

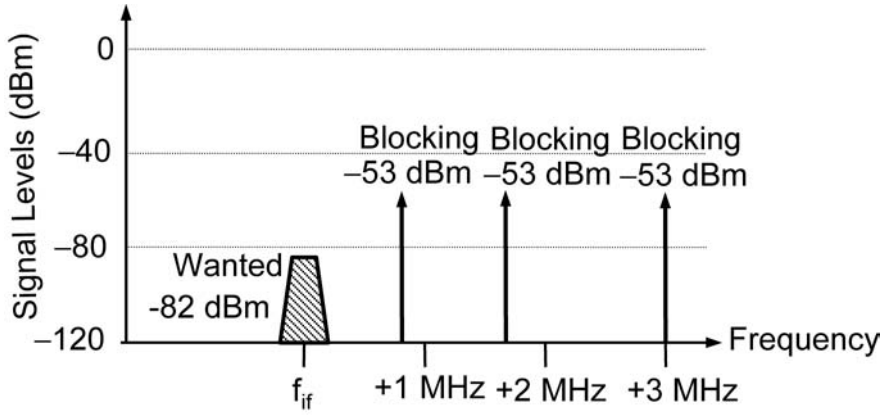

Figure 3.14 Minimum detectable signal and blocker levels after the IF filter.

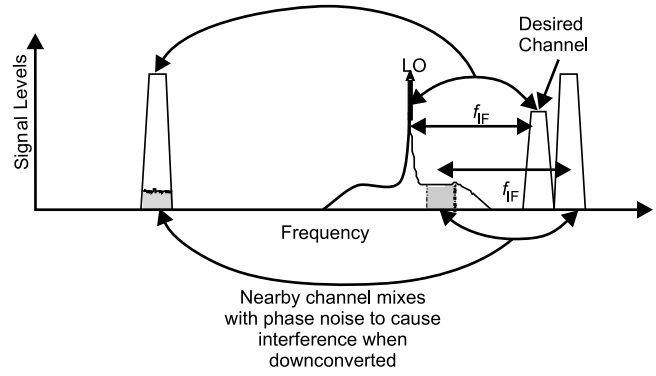

Figure 3.15 Phase noise can cause nearby channels to get mixed down on top of the desired signal.

if the oscillator also has energy at the same offset f from the carrier, then the blocking signal will be mixed directly to the IF frequency as illustrated in Figure 3.15.

Example 3.7: Calculating Maximum Level of Synthesizer Spurs

For the specifications given in the previous example, calculate the allowable noise in a synthesizer in the presence of the blocking signals. Assume the IF bandwidth is 200 kHz.

Solution:
Any tone in the synthesizer at an 800-kHz offset will mix with the blocker, which is at –43 dBm, and mix it to the IF stage where it will interfere with the wanted signal. The blocker can be mixed with noise anywhere in the 200-kHz bandwidth. Since noise is specified as a density, total noise is obtained by multiplying by the bandwidth, or equivalently in dB by adding 10 log 200,000 or 53 dB. We note that to be able to detect the wanted signal reliably, as in the previous example, we need the signal to be about 11 dB or so above the mixed-down blocker. Therefore, the mixed-down blocker must be less than –113 dBm. Therefore, the maximum synthesizer noise power at 800 kHz offset is calculated as –113 + 43 53 = 123 dB lower than the desired oscillating amplitude measured in a 1-Hz bandwidth.

3.2.7 DC Offset

DC offset is caused primarily by leakage of the LO into the signal path. If the LO leaks into the signal path and is then mixed with itself (called self mixing), this mixing product will be at zero frequency as shown in Figure 3.16. This will produce a DC signal or DC offset in the baseband, which has nothing to do with the information in the modulated signal. This offset will be amplified by any gain stages in the baseband and can saturate the ADC if it is too large. DC offsets can also be caused by PA leakage and jamming signals received from the antenna. This problem is often worse in direct conversion radios where there is usually much more gain in the baseband of the radio and the LO is often at a much higher frequency, thus reducing the LO isolation of the radio.

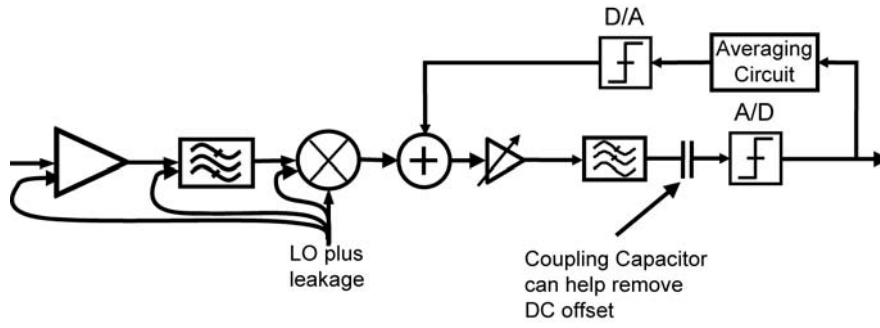

Figure 3.16 LO self-mixing can cause DC offsets.

There are a few things that can be done about DC offset. If the radio uses a modulation type where there is not much information at DC (such as an OFDM signal where the first subcarrier does not contain any information), then a blocking capacitor can be placed right before the ADC. This will act as a highpass filter and will prevent the DC offset from entering the ADC. Since DC offsets are often not time variant it may also be possible to calibrate much of it out of the signal path. This can be done by sensing the DC offset in the baseband and adjusting the output from the mixer to compensate for it.

3.2.8 Second-Order Nonlinearity Issues

Second-order nonlinearity is also very important in transceiver specifications. In this case, the main cause of nonlinearity is the IQ mixer that downconverts the signal to baseband. Consider again the case where there are two in-band interfering signals. If these signals are close to each other in frequency, then the difference between these two frequencies may fall into the baseband bandwidth of the receiver as shown in Figure 3.17. If this happens then the signal-to-distortion ratio must still be large enough to make sure that the signal can be detected with sufficient BER. It is usually only the final downconversion stage that is the problem, as prior to the mixer itself, a simple ac coupling capacitor will easily filter out any low frequency IM2 products produced by the earlier stages in the radio.

Example 3.8: IIP2 Calculation

A direct downconversion receiver is required to detect a signal at 2 GHz with a power level of 80 dBm at the input of the downconversion mixer. An SNR of 15 dB is required and the signal has a bandwidth of 20 MHz. Two interferers are present in band, each with a power level of 20 dBm. They are located at 2,100 MHz and 2,110 MHz. Determine the required IP2 for this system.

Solution:
The second order nonlineaity of the mixer will produce a tone at a frequency of 2,110 2,100 = 10 MHz, which will fall into the band of the desired channel. Since the signal strength is 80 dBm and the SNR required is 15 dB, the power

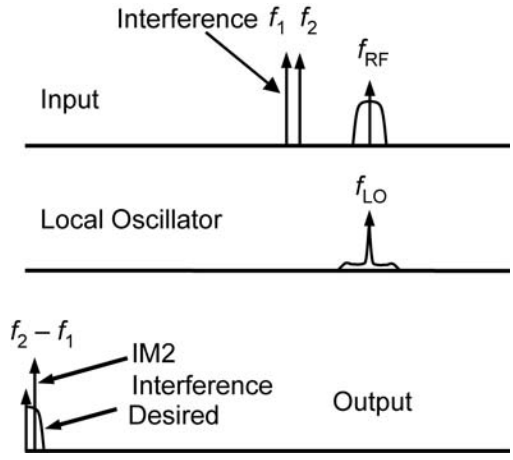

Figure 3.17 Illustration of a second-order product causing distortion.

of this distortion product must be less than 95 dBm. Using (2.66), we note that P2 P1 = (20 (95)) = 75 dB. Therefore the IIP2 must be greater than 20 dBm + 75 dB = 55 dBm.

3.2.9 Receiver Automatic Gain Control Issues

ADCs require relatively constant input amplitude to function properly. They are not equipped to deal with large ranges of input amplitude, and therefore it is one of the most important jobs of the radio to provide them with a constant signal amplitude. On the receive side, as a bare minimum the radio must provide an AGC range at least equal to the dynamic range of the radio. In addition, radio gain will vary with temperature, voltage supply variations, and process. Usually at least an additional 20 dB of gain control is required to overcome these variations. Normally it is possible to use stepped AGC with discrete gain settings (usually spaced by about 3 dB, but maybe less as needed), but some more sophisticated radios may require continuous AGC. It is important to decide where the AGC level will be set and in what order different gain control blocks will be adjusted. At the minimum detectable level, the receiver is set to the maximum gain setting. As the input power starts to rise, it is better to reduce the gain as far towards the back of the radio as possible to have the lowest effect on the noise figure. At high power levels, it is better to adjust the gain at the front of the radio thus reducing the linearity requirements of parts further down the chain.

On the transmit side, AGC is often simpler than on the receive side. Even simple transmitters must have some AGC to compensate for process, temperature, and supply voltage variations. Typically, a simple stepped AGC of about 20 dB is used to make sure that the required power can always be delivered to the antenna. More sophisticated radios will also use power control in the transmitter to back transmit power off if the receiving radio is close by, thus reducing the required maximum received power on the other side of the link.

3.2.10 EVM in Transmitters Including Phase Noise, Linearity, IQ Mismatch, EVM with OFDM Waveforms, and Nonlinearity

Error vector magnitude (EVM) is a very important way to measure how accurately a transmitter has reproduced the vectors that correspond to the data being transmitted. EVM is another way to measure the signal-to-noise and distortion ratios of the signal. The term EVM is used more often in transmitters compared to receivers as a measure of modulation accuracy instead of SNR.

In any system with limited performance, some errors will always be introduced and the vector that is actually transmitted will be different by some amount (an error) from what was intended as shown in Figure 3.18. The instantaneous value of the EVM is defined as the ratio of the magnitude of the reference vector to the magnitude of the error vector

$$EVM_i = \frac{|e_i|}{|a_i|} \tag{3.29}$$

where e_i is the ith error vector and a_i is the ith reference vector. Normally EVM is averaged over a large number of data samples N to come up with an average or nominal value:

$$EVM = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\left(\frac{|e_i|}{|a_i|}\right)^2} \tag{3.30}$$

There are many sources of EVM in a transmitter, and the overall effect of various sources can be added together as shown here:

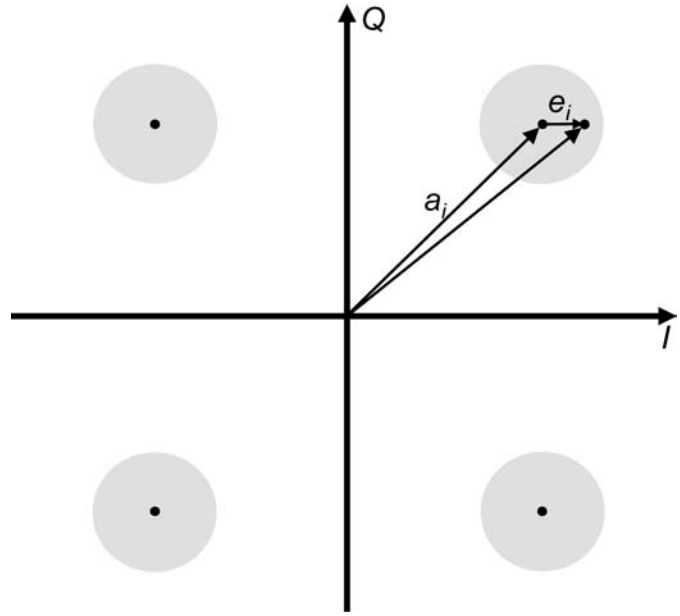

Figure 3.18 Illustration of EVM using QPSK as an example modulation.

$$EVM_{tot} = \sqrt{(EVM_1)^2 + (EVM_2)^2 \ldots + (EVM_M)^2} \tag{3.31}$$

One source of EVM is synthesizer phase noise. As discussed in Chapter 2, the phase noise can be integrated to give a value for the rms phase variation in radians $IntPN_{rms}$. This phase variation will affect the angle of the reference vector and will move it away from its ideal value. Thus, the phase noise will generate an error vector approximately orthogonal to the reference vector as shown in Figure 3.19. In this case the error vector will have a magnitude of:

$$|e_{iPN}| = a_i \sin(IntPN_{rms}) \quad a_i IntPN_{rms} \tag{3.32}$$

Thus the EVM due to phase noise is:

$$EVM_{PN} = \frac{|a_i| \times IntPN_{rms}}{|a_i|} = IntPN_{rms} \tag{3.33}$$

Note that a phase mismatch between the I and Q paths either created by the synthesizer or a mismatch in the baseband paths will behave exactly the same way in regards to EVM as phase noise does. In this case, if there is a phase mismatch in the LO of θ_{LO} then the EVM due to IQ phase mismatch will be:

$$EVM_{LO} = \theta_{LO} \tag{3.34}$$

Another source of EVM is gain mismatch in the baseband I and Q paths leading up to the mixers. If there is some random shift in amplitude from the ideal amplitude in the I and Q paths δ, such that the magnitude of the I path becomes $I\ (1 + \delta)$ and the magnitude of the Q path becomes $Q(1 \quad \delta)$, then the error vector will have a magnitude of:

$$e_i = \sqrt{(\delta I)^2 + (\delta Q)^2} \tag{3.35}$$

The reference vector will have a magnitude of:

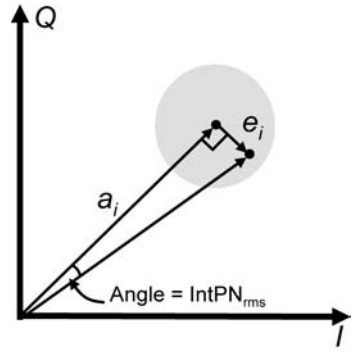

Figure 3.19 Phase noise can cause EVM by rotating the vector.

$$a_i = \sqrt{I^2 + Q^2} \tag{3.36}$$

In this case the EVM will be simply:

$$EVM_{IQ} = \delta \tag{3.37}$$

EVM can also be caused by carrier feedthrough. Carrier feedthrough can be due to finite isolation from LO to RF and from LO to IF, and DC offset, as well as other factors. Whatever the source, any energy from the carrier will distort the waveform and cause EVM. If we define carrier suppression C_s as the ratio of the desired RF power P_t to the LO leakage P_{CFT} then the EVM due to carrier feedthrough is given by:

$$EVM_{CFT} = \sqrt{\frac{P_{\mathrm{CFT}}}{P_t}} = \sqrt{C_s} \tag{3.38}$$

Note that P_{CFT} will be a vector with a random phase relative to the reference vector, but only magnitude is important.

Linearity in the transmitter may also be a consideration in regards to EVM depending on the type of modulation used. If the system uses a phase-only modulation, linearity is of much less concern. With amplitude sensitive modulations like QAM it is best to operate the output with a power level that is well backed off from the 1-dB compression point of the transmitter. Systems that use OFDM have a different set of linearity requirements in regards to EVM than systems that use simpler modulation schemes. If we assume that the transmitter is operating well below the 1-dB compression point and the linearity can be well described by third-order products, the dominant contribution to EVM degradation due to linearity of the transmitter will be triple-order beats of all subcarriers dumping energy on top of the subcarrier of interest. The power in a triple-beat tone as shown in Chapter 2 is given by:

$$TB = OIP_3 \quad 3(OIP_3 \quad P_s) + 6 = 3P_s \quad 2OIP_3 + 6 \tag{3.39}$$

where P_s is the power of the tones generating the triple beat (in this case the power in one subcarrier) and OIP_3 is the output referred third-order intercept point both given in dBm. The EVM can then be computed as the ratio of the power of the subcarrier to the power of all triple beats dumping power into that channel. The difference between the power of the triple beat tones and the desired channel power is therefore:

$$TB \quad P_s = 2P_s \quad 2OIP_3 + 6 + 10\log\left[\frac{3}{8}N^2\right] \tag{3.40}$$

Thus, the EVM due to transmitter nonlinearity is given as:

$$EVM_{lin} = 10^{\frac{2P_s \quad 2OIP_3 + 6 + 10\log\left[\frac{3}{8}N^2\right]}{20}} \tag{3.41}$$

where N is the number of subcarriers. The above equation has been verified through extensive simulations and comparison to measured results by the authors. However, note that if the output power is close to the 1-dB compression point, this formula will become less accurate. Note also that narrowband systems rely on filtering to minimize the number of tones that can intermodulate, making TB tones much less of an issue. However, OFDM uses multiple subcarriers that are always present; hence, it is much more like a video system in which a whole range of tones can intermodulate, even though EVM or BER, but not CTB is typically used as a measure of performance.

If the transmitter is backed off and transmitting at a low power level then the noise in the transmitter circuits can become an important contribution to the EVM of the system. If the total noise power density coming from the transmitter is N_o then the EVM contribution from noise is given by:

$$EVM_{noise} = \sqrt{\frac{N_o \times BW}{P_t}} \tag{3.42}$$

where BW is the bandwidth of the channel.

3.2.11 ADC and DAC Specifications

While this book will focus on RFICs in detail, it is important to say at least a few words about the analog-to-digital converters (ADC) and digital-to-analog converters (DAC), as these blocks are the interface between the radio and the DSP. Generally from a system perspective it is important to specify both the number of bits and the jitter or phase noise requirement of the reference clock, which drives the converters. Both these properties will affect the BER by influencing the SNR.

Suppose that a quantizer converts a continuous analog signal to a discrete digital signal with a characteristic shown in Figure 3.20, where x is the analog input

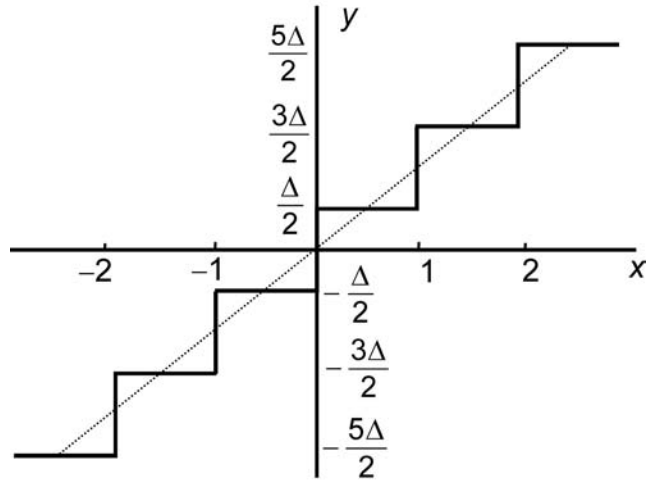

Figure 3.20 Transfer characteristic of a multibit quantizer.

and y is the quantized digital output. The output is a function of the input, but it has discrete levels at equally spaced intervals . Thus, unless the input happens to be an integer multiple of the quantizer resolution (step size) , there will be an error in representing the input. This error, e, will be bounded over one quantizer level by a value of:

$$\frac{}{2} \quad e \quad \frac{}{2} \tag{3.43}$$

Thus, the quantized signal y can be represented by a linear function with an error e as

$$y = \quad \times x + e \tag{3.44}$$

where the step size corresponds to the slope of the straight line shown in Figure 3.20. The quantization error as a function of the input is given in Figure 3.21. Note that the error is a straight line with a slope of – . If the input is random in nature then the instantaneous error will also be random. The error is thus uncorrelated from sample to sample and can hence be treated as noise. Quantization and the resultant quantization noise can be modeled as a linear circuit including an additive quantization error source as shown in Figure 3.22.

Thus, the quantization noise for a random signal can be treated as additive white noise having a value anywhere in the range from – /2 to /2. The quantization noise has a probability density of:

$$p(e) = \begin{cases} \dfrac{1}{} & \text{if} \quad \dfrac{}{2} \quad e \quad \dfrac{}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3.45}$$

where the normalization factor 1/ is needed to guarantee that:

$$\int_{/2}^{/2} p(e)de = 1 \tag{3.46}$$

The mean square rms error voltage e_{rms} can be found by integrating the square of the error voltage and dividing by the quantization step size,

$$e_{rms}^2 = \int^{+} p(e)e^2 de = \frac{1}{} \int_{/2}^{+/2} e^2 de = \frac{^2}{12} \tag{3.47}$$

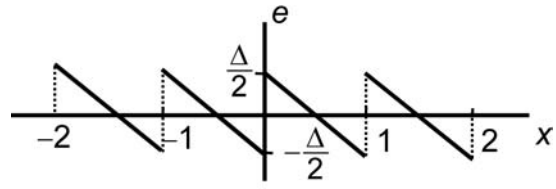

Figure 3.21 The quantization error as a function of the input.

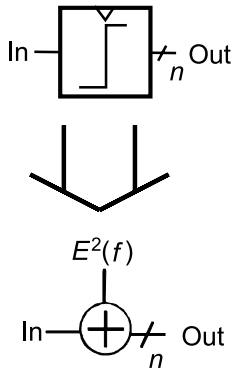

Figure 3.22 Modeling of the quantization as a linear circuit with an additive noise.

From sampling theory, it is known that the frequency spectrum of a sampled system repeats once every sampling frequency. Thus, the spectrum of the quantization noise in a sampled system will be centered around dc and spread out to half of the sampling frequency $f_s/2$ and there will be a copy of the noise spectrum from $f_s/2$ to $3f_s/2$ and so on. Considering that all the noise power lies in the range of positive frequency band (i.e., 0 f) the quantization noise power thus folds into the band from dc to $f_s/2$. Assuming white noise, the power spectral density $E^2(f)$ of the quantization noise is given by:

$$E^2(f) = \frac{e^2_{rms}}{f_s/2} = 2Te^2_{rms} \tag{3.48}$$

where the sample period $T = 1/f_s$. For a band limited signal 0 $f < f_0$ with bandwidth of f_0, the quantization noise power that falls into the signal band can be found as:

$$n_0^2 = \int_0^{f_0} E^2(f)df = 2f_oTe^2_{rms} = \frac{^2f_0}{6\times f_s} = \frac{^2}{12\times \text{OSR}} \tag{3.49}$$

where the oversampling rate (OSR) is defined as ratio of the sampling frequency f_s to the Nyquist frequency $2f_0$, that is:

$$\text{OSR} = \frac{f_s}{2f_0} \tag{3.50}$$

In an N-bit sampled system, if the quantizer has $2N$ quantization levels equally spaced by , then the maximum peak-to-peak amplitude is given by:

$$\upsilon_{\max} = (2^N \quad 1)\times \tag{3.51}$$

If the signal is sinusoidal, the associated signal power is:

$$P = \frac{1}{8}(2^N \quad 1)^2 \times \ ^2 \tag{3.52}$$

Thus, the signal-to-noise ratio (SNR) due to quantization noise power that falls into the signal band becomes:

$$\mathrm{SNR} = 10\log\left(\frac{\frac{1}{8}\left(2^N - 1\right)^2}{n_0^2}\right) - 10\log\left(\frac{3\cdot 2^{2N}\mathrm{OSR}}{2}\right) \tag{3.53}$$

Noting that $\log_{10}(x) = \log_{10}(2) \times \log_2(x)$, the above expression becomes:

$$\mathrm{SNR} \approx 6.02 \times N + 3 \times \log_2(\mathrm{OSR}) + 1.76 \tag{3.54}$$

Therefore, the SNR improves by 6 dB for every bit added to the quantizer. For the same amount of total quantization noise power, every doubling of the sampling frequency reduces the in-band quantization noise by 3 dB.

The other main source of noise in an ADC is the timing jitter of the reference clock. Timing jitter t_{jitter} is related to the integrated phase noise of the of the reference clock by:

$$t_{jitter} = \frac{IntPN_{rms}}{2\pi} \times T_{clk} \tag{3.55}$$

where T_{clk} is the period of the reference clock. Now if the input to the ADC is assumed to be a sine wave at a frequency f_{in} the amount of noise caused by reference clock jitter can be estimated. If the input waveform is given by:

$$v_{in}(t) = A\sin(2\pi \times f_{in} t) \tag{3.56}$$

then the slope of the input is simply:

$$\frac{dv_{in}(t)}{dt} = 2\pi \times f_{in} A\cos(2\pi \times f_{in} t) \tag{3.57}$$

This slope will be highest at the waveform zero crossings and if there is an error in the time the waveform is sampled t_{jitter} then there will be an rms error in the sampled voltage of:

$$v_{in\_error\_rms} = 2\pi \times \frac{f_{in} A}{\sqrt{2}} \times t_{jitter} \tag{3.58}$$

In this case the SNR will be given by:

$$SNR_{jitter} = -20\log(2\pi \times f_{in} \times t_{jitter}) \tag{3.59}$$

Therefore it is now possible to estimate the number of bits and clock jitter required to assure that the SNR of the ADC or DAC does not degrade the overall system SNR by more than the allowed amount.

Example 3.9: Specifying an ADC

An ADC is required to have a SNR of 30 dB and a signal bandwidth of 20 MHz. Ignoring any out of band interferers, give as much detail as possible about the performance the ADC would require.

Solution:
First let us assume that we are going to use a Nyquist rate ADC for this design and therefore the ADC will be clocked at 40 MHz and the OSR will be one. To give some margin in the design we will calculate the number of bits required for a SNR of 33 dB. In this case making use of (3.54), this ADC will need:

$$N = \frac{SNR \quad 1.76}{6.02} = 5.2$$

Thus, a 6-bit ADC will be adequate. Next we can calculate the requirement on the reference clock jitter using (3.59). Again assuming 33 dB to give some margin:

$$t_{jitter} = \frac{\log^{1} \quad \frac{SNR_{jitter}}{20} \div}{2\pi \times f_{in}} = 0.18 \text{ ns}$$

Therefore, this application will require a reference clock with better than 0.18 ns of jitter.

3.3 Antennas and the Link Between a Transmitter and a Receiver

In general, information can be transmitted through cables, fiber, and other media. Transmitters and receivers designed with RFIC technology most often use air as the channel link between the transmitter and the receiver. In this case, there must be an antenna attached to both radios to transform between RF signals in the electronic circuits and EM radiation in the air. The antenna attached to the transmitter radiates the EM signal into the air. It is then the job of the receiving antenna to collect some fraction of the energy transmitted and provide this energy to the receiver so that the signal can be detected and processed. The design of antennas of all descriptions could take up a whole book by itself, however some very basic information will be given here in order to understand some of the limitations of transmitting information across a radio link. Antennas can either be isotropic, which means that energy radiates equally in all directions, or alternatively the antenna can be directional, which focuses more energy in a particular direction in order to maximize the amount of energy that can be detected across the link. The amount of directivity that an antenna possesses can be thought of as an antenna gain. The antenna gain refers to the amount of excess energy transmitted in a given direction over and above that which would have been transmitted by an isotropic antenna. It can be shown that in free space the amount of power collected by a receiving antenna is given by [9]:

$$P_{\mathrm{r}} = \frac{P_t G_t G_r \lambda^2}{(4\pi d)^2} \tag{3.60}$$

where P_t is the power transmitted by the transmitting antenna, G_t is the gain of the transmitting antenna, G_r is the gain of the receiving antenna, λ is the wavelength, and d is the distance separating the two antennas.

If it is assumed that there is both a direct line of sight path and a path of energy reflected from the ground, then the formula must be modified. If the transmitter is at a height of h_t and the receiver is at a height of h_r, as shown in Figure 3.23, then the received power is given by [9]:

$$P_{\mathrm{r}} = \frac{P_t G_t G_r h_t^2 h_r^2}{d^4} \tag{3.61}$$

provided that:

$$d >> \sqrt{h_t h_r} \tag{3.62}$$

This problem becomes more complicated if there is an obstruction in the way between the transmitter and the receiver. If the line of site path is blocked then it may still be possible to receive some signal in the "shadow" of the obstruction; the EM waves will bend around the object, but the signal strength will be reduced. For a half plain obstruction sometimes called a knife edge obstruction, as shown in Figure 3.24, the magnitude of the reduction in the received signal power relative to that of free space will be given by [9]:

$$G_d = 20\log|F(\upsilon)| \tag{3.63}$$

where

$$F(\upsilon) = \frac{(1+j)}{2} \int_{t=\upsilon}^{t=} \exp\left(\frac{j\pi t^2}{2}\right) dt = \frac{E_d}{E_o} \tag{3.64}$$

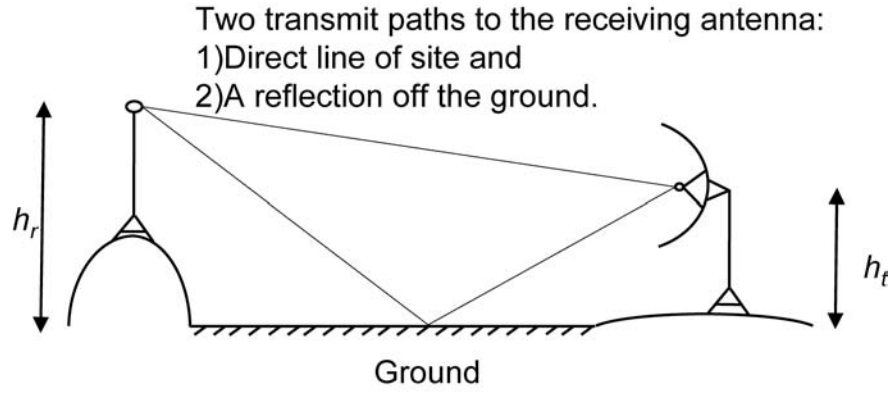

Figure 3.23 Two fixed antennas with a ground plane.

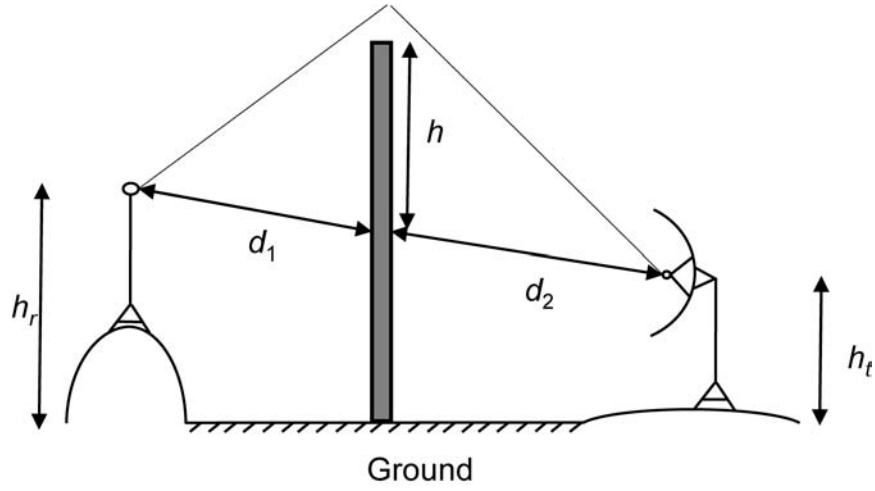

Figure 3.24 Two antennas with a knife edge obstruction in the way.

Note that $F(\upsilon)$ is the ratio of the free space EM field strength to the actual received field strength and,

$$\upsilon = h\sqrt{\frac{2(d_1 + d_2)}{\lambda d_1 d_2}} \tag{3.65}$$

Obviously real environments are far more complicated then this and a theoretical formula cannot be derived for all practical situations, but a modified version of (3.60) can be used to provide useful estimates for path loss in a number of common environments. The average path loss can be estimated as

$$PL = PL_{ref}(d_o) + 10n\log\left[\frac{d}{d_o}\right] \tag{3.66}$$

where d_o is the reference distance at which point the reference path loss PL_{ref} is calculated, and n is the path loss exponent. Some common values for n are given in Table 3.1. Note two different points at the same distance from the transmitter could have different actual path losses. The reference distance must be chosen to be in the far field of the transmitting antenna. Typically for long distance outdoor applications this might be of the order of 1 km, while for indoor applications a distance

Table 3.1 Path Loss Exponent in Various Environments

Environment	*Path Loss Exponent n*
Free space	2
Urban cellular radio	2.7 to 3.5
Shadowed urban cellular radio	3 to 5
In building line of sight	1.6 to 1.8
Obstructed in buildings	4 to 6
Obstructed in factories	2 to 3

of the order of 1m might be used. The reference path loss $PL_{ref}(d_o)$ can either be calculated from the free space loss (3.60) or it may be measured.

Example 3.10: Estimating Received Power Levels

For a 5-GHz indoor link with transmit and receive antenna gain of 6 dBi, what transmit power is required to ensure that 80 dBm is received at a distance of 30m?

Solution: First we will calculate the reference path loss at 1m from the transmitter using the free space equation (3.60) and noting that a wavelength at 5 GHz is 0.06m:

$$PL_{ref}(d_o) = \frac{(4\pi d_o)^2}{G_t G_r \lambda^2} = \frac{(4\pi \times 1)^2}{4 \times 4 \times 0.06^2} = \frac{157.75}{0.576} = 273.9 = 24.4 \text{ dB}$$

Now, referring to Table 3.1, for an indoor obstructed environment we will use a path loss exponent of $n = 5$. Thus the total path loss over a distance of 30m will be:

$$PL = PL(d_o) + 10n \log \left(\frac{d}{d_o} \right) = 24.4 + 10 \times 5 \times \log \left(\frac{30}{1} \right) = 98.3 \text{ dB}$$

Thus to receive 80 dBm at a distance of 30m means that a transmit power of 18.3 dBm is required.

There is much more that could be said about communication links. The interested reader is referred to [9] for more information.

References

[1] Razavi, B., *RF Microelectronics*, Upper Saddle River, NJ: Prentice-Hall, 1998.

[2] Crols, J., and M. Steyaert, *CMOS Wireless Transceiver Design*, Dordrecht, the Netherlands: Kluwer Academic Publishers, 1997.

[3] Couch II, L. W., *Digital and Analog Communication Systems*, 6th ed., Upper Saddle River, NJ: Prentice Hall, 2001.

[4] Carson, R.S., *Radio Communications Concepts: Analog*, New York: John Wiley & Sons, 1990.

[5] Rohde, U. L., J. Whitaker, and A. Bateman, *Communications Receivers: DSP, Software Radios, and Design*, 3rd ed., New York: McGraw-Hill, 2000.

[6] Larson, L. E., (ed.), *RF and Microwave Circuit Design for Wireless Communications*, Norwood, MA: Artech House, 1997.

[7] Gu, Q., RF *System Design of Transceivers for Wireless Communications*, 1st ed., New York: Springer, 2005.

[8] Sheng, W., A. Emira, and E. Sanchez-Sinencio, "CMOS RF Receiver System Design: A Systematic Approach," *IEEE Trans. on Circuits and Systems I: Regular Papers*, Vol. 53, No. 5, May 2006, pp. 1023–1034.

[9] Rappaport, T. S., *Wireless Communications: Principles and Practice*, 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 2001.

CHAPTER 4

A Brief Review of Technology

4.1 Introduction

At the heart of RF integrated circuits are the transistors used to build them. The basic function of a transistor is to provide gain. Unfortunately, transistors are never ideal because along with gain comes nonlinearity and noise. The nonlinearity is used to good effect in mixers and in the limiting function in oscillators. Transistors also have a maximum operating frequency beyond which they cannot produce gain.

Metal oxide semiconductor (MOS) and bipolar transistors will be discussed in this chapter. CMOS is the technology of choice in any digital application because of its very low quiescent power dissipation and ease of device isolation. However, traditionally MOS field-effect transistors (MOSFETs) have had inferior speed and noise compared to bipolar transistors. Also, CMOS devices have proved challenging to model for RF circuit simulation, and without good models, RFIC design can be a very frustrating experience. In order to design RFICs, it is necessary to have a good understanding of the high-speed operation of the transistors in the technology that is being used. Thus, in this chapter a basic introduction to some of the more important properties will be provided. For more detail on transistors, the interested reader should consult [1–10].

4.2 Bipolar Transistor Description

Figure 4.1 shows a cross section of a basic npn bipolar transistor. The collector is formed by epitaxial growth in a p substrate (the n region). A p region inside the collector region forms the base region; then an n+ emitter region is formed inside the base region. The basic transistor action all takes place directly under the emitter in the region shown with an oval. This can be called the intrinsic transistor. The intrinsic transistor is connected through the diffusion regions to the external contacts labeled *e*, *b*, and *c*. More details on advanced bipolar structures, such as using SiGe heterojunction bipolar transistors (HBTs), and double-poly self-aligned processes can be found in the literature [1, 2]. Note that although Si is the most common substrate for bipolar transistors, it is not the only one; for example, GaAs HBTs are often used in the design of cellular radio power amplifiers and other high power amplifiers.

Figure 4.2 shows the transistor symbol and biasing sources. When the transistor is being used as an amplifying device, the base-emitter junction is forward biased while the collector-base junction is reverse biased, meaning the collector is

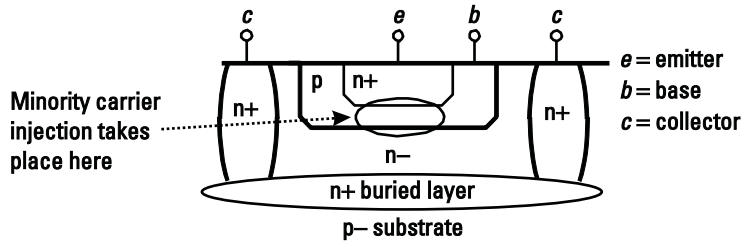

Figure 4.1 Planar bipolar transistor cross-section diagram.

at a higher voltage than the base. This bias regime is known as the forward active region. Electrons are injected from the emitter into the base region. Because the base region is narrow, most electrons are swept into the collector rather than going to the base contact. This is equivalent to conventional (positive) current from collector to emitter. Some holes are back injected into the emitter and some electrons recombine in the base, resulting in a small base current, which is directly proportional to collector current $i_c = \beta i_b$. Thus, the overall concept is that collector current is controlled by a small base current. The collector current can also be related to the base-emitter voltage in this region of operation by:

$$I_c = I_s e^{\frac{V_{BE}}{v_t}} \tag{4.1}$$

where I_S is a constant known as the saturation current, V_{BE} is the dc bias between the base and emitter and v_T is the thermal voltage given by:

$$v_T = \frac{kT}{q} \tag{4.2}$$

where q is the electron charge, T is the temperature in Kelvin, and k is Boltzmann's constant. The thermal voltage is approximately equal to 25 mV at a temperature of 290K, close to room temperature.

Figure 4.3 shows the collector characteristics for a typical bipolar transistor. The transistor has two other regions of operation usually avoided in analog design. When the base-emitter junction is not forward biased, the transistor is cut off. The transistor is in the saturated region if both the base-emitter and collector-emitter junction are forward biased. In saturation, the base is flooded with minority carriers. This generally leads to a delayed response when the bias conditions change to

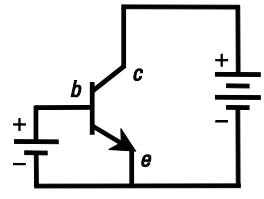

Figure 4.2 Bipolar transistor symbol and bias supplies.

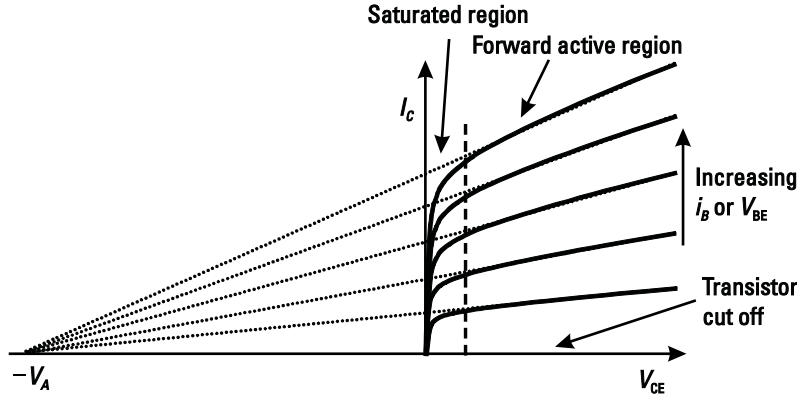

Figure 4.3 Transistor characteristic curves.

another region of operation. In saturation, V_{CE} is typically less than a few tenths of a volt. Note that in the active region, the collector current is not constant. There is a slope to the current versus voltage curve, indicating that the collector current will increase with collector-emitter voltage. The slopes of all the lines are such that they will meet at a negative voltage V_A called the *Early voltage*. This voltage can be used to characterize the transistor output impedance.

The intrinsic transistor is connected through the diffusion regions to the external contacts labeled *e*, *b*, and *c*. These connections add series resistance and increase the parasitic capacitance between the regions. The series resistance in the collector is reduced by the buried layer. The effects of other series resistance are often reduced by the use of multiple contacts, as shown in Figure 4.4.

The maximum allowed voltage is typically limited by the breakdown voltage with specifications usually found in the process documentation. Maximum current is dependent on the transistor dimensions and layout. Because bipolar transistors can suffer from localized hot spots and so-called *secondary breakdown*, at higher voltages the maximum allowable current is reduced. This is sometimes described as the safe operating area (SOA). Again, process documentation should be consulted for specific numbers. Since these topics are of most importance in power amplifiers, further information is given in Chapter 11.

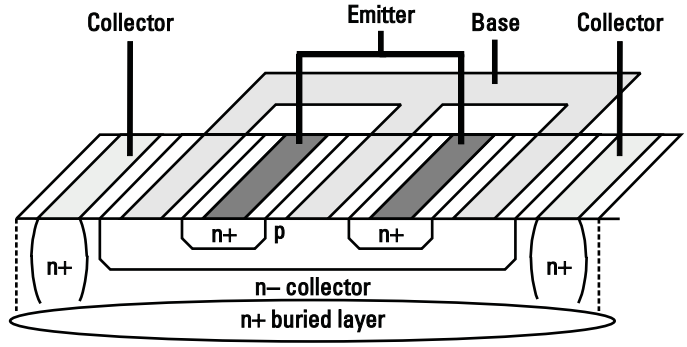

Figure 4.4 Transistor with multiple contacts, shown in three dimensions.

Figure 4.5 Current dependency of β.

4.3 β Current Dependence

Figure 4.5 shows the dependence of β on the collector current. β drops off at high currents because the electron concentration in the base-collector depletion region becomes comparable to the background dopant ion concentration, leading to a dramatic increase in the effective width of the base. This is called the *Kirk effect* or *base pushout*. As a result, the base resistance is current dependent. Another effect is emitter crowding, which comes about because of the distributed nature of parasitic resistance in the base region, causing the junction forward bias voltage to be higher at the emitter periphery closest to the base contact. This results in the highest current density at the edge of the emitter. In the other extreme, at low currents, β may be reduced due to the excess current resulting from recombination in the E-B depletion region.

4.4 Small-Signal Model

Once the bias voltages and currents are determined for the transistor, it is necessary to determine how it will respond to alternating current (ac) signals exciting it. Thus, an ac small-signal model of the transistor is now presented. Figure 4.6 shows a fairly complete small-signal model for the bipolar transistor. The values of the small-signal elements shown, r_π, C_π, C_μ, g_m, and r_o will depend on the dc bias of the transistor. The intrinsic transistor (shown directly under the emitter region in Fig-

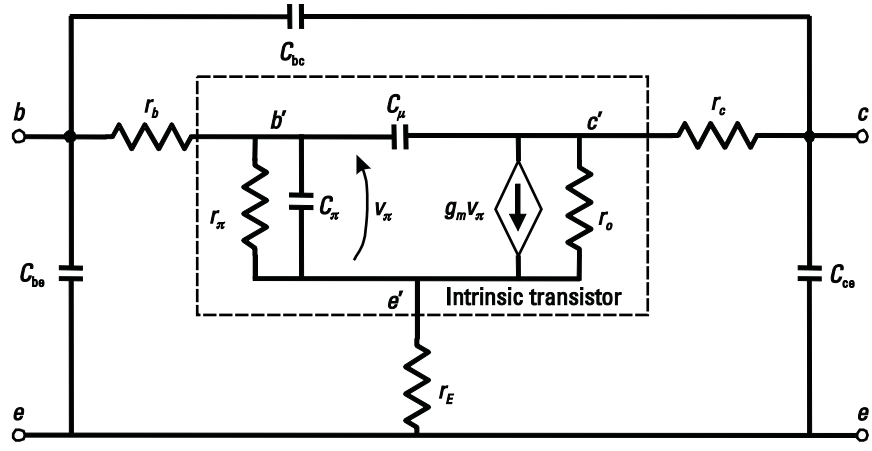

Figure 4.6 Small-signal model for bipolar transistor.

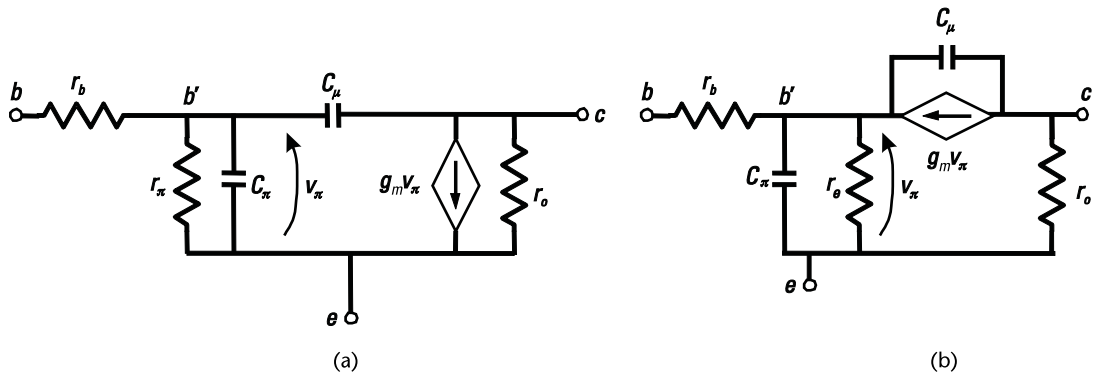

Figure 4.7 Simplified small-signal models for bipolar transistor: (a) pi-model and (b) T-model.

ure 4.1) is shown at the center. The series resistances to the base, emitter and collector are shown respectively by r_b, r_E, and r_c. Also, between each pair of terminals there is some finite capacitance shown as C_{bc}, C_{ce}, C_{be}. This circuit can be simplified by noting that of the extrinsic resistors, r_b is the largest and as a result, r_E and r_c are often omitted along with the capacitances C_{bc}, C_{ce}, C_{be}. Resistor r_E is low due to high doping of the emitter while r_c is reduced by a buried layer in the n-collector. The base resistance r_b is the source of several problems. First, it forms an input voltage divider between r_b, r_π, and C_π, which reduces the input signal amplitude and deteriorates high-frequency response. It also directly adds to thermal noise. The resulting simplified small-signal models are shown as a pi-model in Figure 4.7(a) and as a T-model as in Figure 4.7(b). While r_π is associated with the pi-model and r_e is associated with the T-model, it can be easily verified, for example with SPICE simulations, that both models have identical terminal impedances and identical frequency responses for all values of driving or load impedances. Note that at low frequencies, r_π has base current flowing through it while r_e has emitter current flowing through it, hence for v_π to be the same in both models, r_e is smaller than r_π by a factor of $(\beta + 1)$. We note that with either model, we may use r_e as an estimate of resistance seen looking into the emitter (the estimate is exact at low frequencies if r_b is small and the base is grounded). We also note that r_e is approximately equal to the inverse of g_m and this will be used extensively in various parts of the book.

4.5 Small-Signal Parameters

Now that the small-signal model has been presented, some simple formulas will be presented to help determine what model parameter values should be at different operating points.

First, the short-circuit current gain β is given by

$$\beta = \underbrace{\frac{i_c}{i_b}}_{\text{small-signal}} = \underbrace{\frac{\Delta I_C}{\Delta I_B}}_{\Delta\text{large-signal}} \tag{4.3}$$

noting that currents can be related by:

$$i_c + i_b = i_e \tag{4.4}$$

Transconductance g_m is given by:

$$g_m = \frac{i_c}{v_\pi} = \frac{I_C}{v_T} = \frac{I_c q}{kT} \tag{4.5}$$

where I_C is the dc collector current. Note that the small-signal value of g_m in (4.5) is related to the large-signal behavior of (4.1) by differentiation.

At a low frequency where the transistor input impedance is resistive, i_c and i_b can be related by:

$$i_c = \beta i_b = g_m v_\pi = g_m i_b r_\pi \tag{4.6}$$

(neglecting current through r_o) which means that:

$$\beta = g_m r_\pi \tag{4.7}$$

As well, the output resistance can be determined in terms of the early voltage V_A:

$$r_o = \frac{V_A}{I_C} \tag{4.8}$$

4.6 High-Frequency Effects

There are two typical figures of merit f_T and $f_{\max}$ used to describe how fast a transistor will operate. f_T is the frequency at which the short-circuit current gain β is equal to 1. $f_{\max}$ is the frequency at which the maximum available power gain $G_{A,\max}$ is equal to 1.

Referring to Figure 4.7, an expression can be found for the corner frequency f_β, beyond which the current gain β decreases:

$$f_\beta = \frac{1}{2\pi r_\pi (C_\pi + C_\mu)} \quad \frac{1}{2\pi r_\pi C_\pi} \tag{4.9}$$

Since this is a first-order roll-off, f_T is β times higher than f_β.

$$f_T = \beta_0 f_\beta = \frac{g_m}{2\pi (C_\pi + C_\mu)} \quad \frac{g_m}{2\pi C_\pi} = \frac{I_C}{2\pi C_\pi v_T} \tag{4.10}$$

The maximum frequency for which power gain can be achieved is called $f_{\max}$, while $G_{A,\max}$ is the maximum achievable gain at a particular frequency. $f_{\max}$ and $G_{A,\max}$ are measured by conjugately matching the source and the load to the transistor.

$f_{\max}$ can be determined by noting that at $f_{\max}$ the impedance of C_π is very low, and the presence of C_μ and the Miller effect makes the impedance even lower. As a result, r_π can be ignored and the input impedance is approximately equal to r_b (the

residual capacitive reactance can be resonated with a series inductor so it can be ignored). Thus input power is:

$$P_{\text{in}} = \frac{v_b^2}{r_b} \tag{4.11}$$

where v_b is the input rms voltage on the base. The current source has an output current equal to

$$i_c = g_m v_\pi \tag{4.12}$$

where the magnitude of v_π is given by

$$|v_\pi| = \frac{v_b}{r_b \omega k C_\pi} \tag{4.13}$$

since the current in C_π is much greater than in r_π. Here k is the multiplier due to the Miller multiplication of C_μ [3]. This factor is often ignored in literature, but it will be shown later that at $f_{\max}$ $k = 3/2$ so should not be ignored. Thus, the output current is

$$i_c = \frac{g_m v_b}{r_b \omega k C_\pi} \tag{4.14}$$

The output impedance is determined by applying an output test voltage v_{cx} and measuring the total output current i_{cx}. The most important component of current comes from the current source $i_{cx} = g_m v_{\pi x}$ where $v_{\pi x}$ is related to v_{cx} by the C_μ, C voltage divider described by

$$v_{\text{x}} = \frac{v_{cx} C_\mu}{C_\pi + C_\mu} \tag{4.15}$$

Note that $r_b + R_S$ has been ignored since its impedance is significantly larger than that of C at $f_{\max}$. Thus the real part of the output impedance z_o is

$$\{z_0\} = \frac{v_{cx}}{i_{\text{cx}}} = \frac{C_\pi + C_\mu}{g_m C_\mu} \quad \frac{C_\pi}{g_m C_\mu} \tag{4.16}$$

Current through C_μ will be seen as a reactive part to the output impedance. It turns out that the real and reactive components are roughly equal; however, the reactive component can be ignored, since its effect will be eliminated by output matching. The remaining real component will be loaded with an equal real component resulting in a voltage ratio of:

$$\frac{v_c}{v_\pi} = \frac{C_\pi}{2C_\mu} \tag{4.17}$$

Using this result, the Miller multiplication of C_μ results in

$$kC_\pi = C_\pi + \quad 1 \quad \frac{v_c}{v_\pi} \div C_\mu = C_\pi + \quad 1 + \frac{C_\pi}{2C_\mu} \div C_\mu = C_\pi + C_\mu + \frac{C_\pi}{2} \quad \frac{3}{2} C_\pi \tag{4.18}$$

Because the output is matched, it is assumed that half the current flows into the output impedance and half the current flows into the load and thus the real power into the load is $i^2R/4$. Then, with the help of (4.14) and (4.16), the output power P_o is:

$$P_o = \frac{i_c^2 \quad \{z_0\}}{4} = \frac{g_m v_b^2}{4r_b^2 \omega^2 k^2 C_\pi C_\mu} \tag{4.19}$$

$$\frac{P_o}{P_i} = \frac{g_m}{4r_b \omega^2 k^2 C_\pi C_\mu} \tag{4.20}$$

If this is set equal to 1, one can solve for $f_{\max}$.

$$f_{\max} = \frac{1}{2\pi} \sqrt{\frac{g_m}{4r_b k^2 C_\pi C_\mu}} = \sqrt{\frac{f_T}{8\pi r_b k^2 C_\mu}} = \frac{1}{4\pi r_{bb}} \sqrt{\frac{\beta_0}{k^2 C_\pi C_\mu}} \tag{4.21}$$

where r_{bb} is the total base resistance given by $r_{bb} = \sqrt{r_b r_\pi}$. Note that $f_{\max}$ can be related to the geometric mean of f_T and the corner frequency defined by r_b and C_μ.

4.6.1 f_T as a Function of Current

f_T is heavily bias dependent, therefore only when properly biased at a current of $I_{\text{opt}f_T}$ will the transistor have its maximum f_T as shown in Figure 4.8. As seen in (4.10), f_T is dependent on C_π and g_m. The capacitor C_π is often described as being a combination of the base-emitter junction capacitance C_{je} and the diffusion capacitance C_d. The junction capacitance is voltage dependent, where the capacitance decreases at higher voltage. The diffusion capacitance is current dependent and increases with increasing current. However, for current levels below the current for peak f_T, g_m increases faster with increasing current; hence, f_T is increasing in this region. At high currents, f_T drops due to current crowding and conductivity modulation effects in the base region [1].

Note that in many processes, f_T is nearly independent of size for the same current density in the emitter (but always a strong function of current). Some f_T curves that could be for a typical modern 50-GHz SiGe process are shown in Figure 4.9.

Example 4.1: f_T and $f_{\max}$ Calculations

From the data in Table 4.1 for a typical 50-GHz bipolar process, calculate z_o, f_T, and $f_{\max}$ for the 15 transistor. Use this to verify some of the approximations made in the above derivation for $f_{\max}$.

Solution:

At 7.9 mA, g_m is equal to 316 mA/V and if $\beta = 100$, then $r_\pi = 316.5$ and f_T is calculated to be 71.8 GHz. It can be noted that a simulation of the complete model

Table 4.1 Example Transistors

	Transistor Size		
Parameter	*1×*	*4×*	*15×*
$I_{opt f_T}$ (mA)	0.55	2.4	7.9
C_π (fF)	50	200	700
C_μ (fF)	2.72	6.96	23.2
r_b ( )	65	20.8	5.0

resulted in a somewhat reduced value of 60 GHz. At 71.8 GHz, the impedance of C_π is calculated to be $j3.167$. Thus, the approximation that this impedance is much less than r_b or r_π is justified. Calculation of $f_{\max}$ results in a value of 104.7 GHz. The real part of the output impedance is calculated as 98.6 . We note that the reactive part of the output impedance will be cancelled out by the matching network.

4.7 Noise in Bipolar Transistors

In addition to the thermal noise in resistors, as discussed in Chapter 2, transistors also have other types of noise. These will be discussed next.

4.7.1 Thermal Noise in Transistor Components

The components in a transistor that have thermal noise are r_b, r_E, and r_c. Given a resistor value R, a noise voltage source must be added to the transistor model of value $4kTR$, as discussed in Chapter 2.

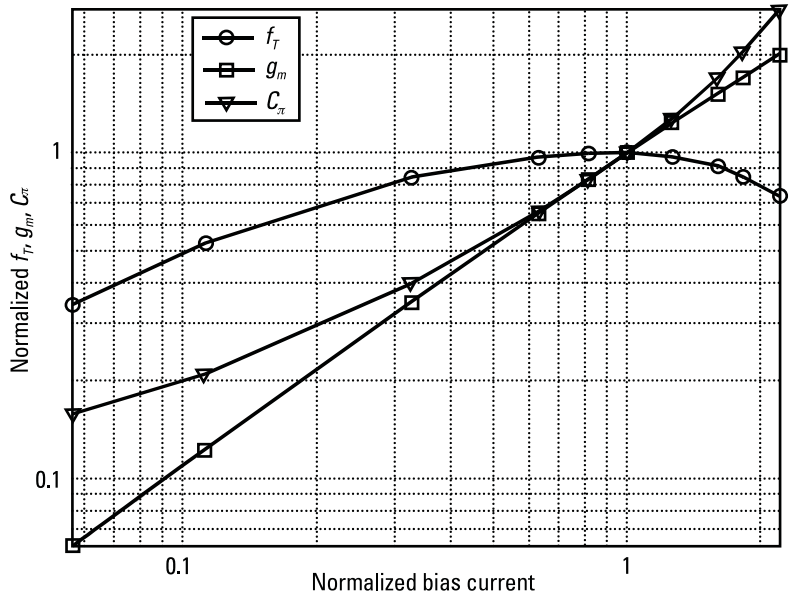

Figure 4.8 Normalized f_T, g_m, and C versus bias current. Normalization is with respect to the values at the optimal f_T point.

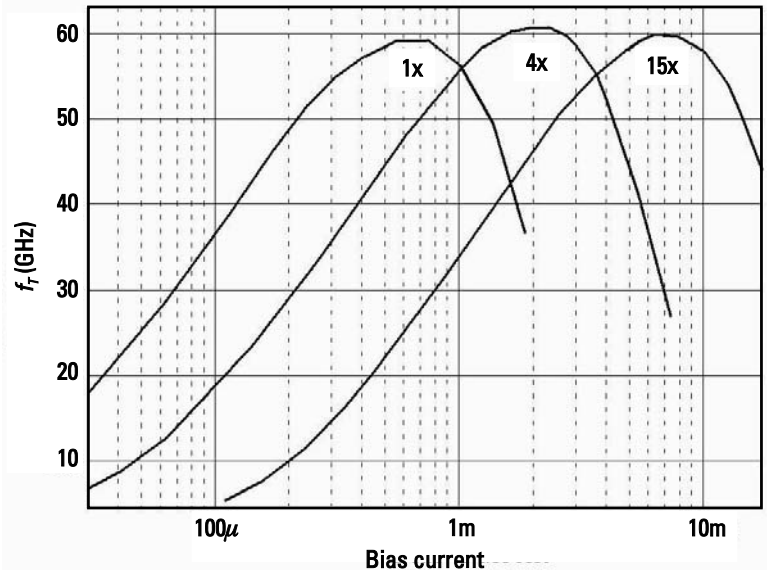

Figure 4.9 f_T as a function of currents for different transistor size relative to a unit transistor size of 1x.

4.7.2 Shot Noise

Shot noise occurs at both the base and the collector, and is due to the discrete nature of charge carriers as they pass a potential barrier, such as a pn junction. That is to say that even though we think about current as a continuous flow, it is actually made up of many electrons (charge carriers) that move through the conductor. If the electrons encounter a barrier they must cross, then at any given instant a different number of electrons will cross that barrier even though on average they cross at the rate of the current flow. This random process is called *shot noise* and is usually expressed in amperes per root hertz. The base shot noise is described by:

$$i_{\text{bn}} = \sqrt{2qI_B} \tag{4.22}$$

and the collector shot noise is described by:

$$i_{\text{cn}} = \sqrt{2qI_C} \tag{4.23}$$

where I_B and I_C are the base and collector bias currents, respectively. The frequency spectrum of shot noise is white.

4.7.3 1/*f* Noise

This type of noise is also called flicker noise, or excess noise. 1/*f* noise is due to variation in the conduction mechanism, for example fluctuations of surface effects (such as the filling and emptying of traps) and of recombination and generation mechanisms. Typically, the power spectral density of 1/*f* noise is inversely proportional to the frequency and is given by the following equation:

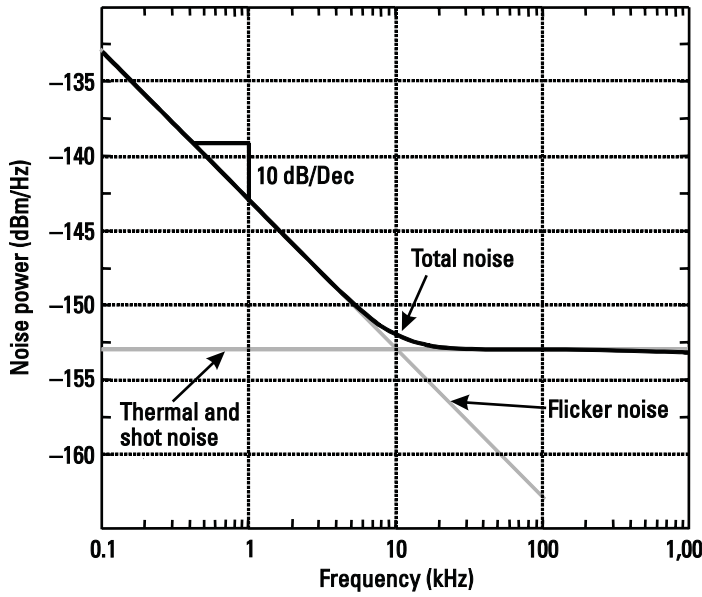

Figure 4.10 Illustration of noise power spectral density.

$$i_{\text{bf}}^2 = KI_C^m \frac{1}{f^\alpha} \tag{4.24}$$

where m is between 0.5 and 2, α is about equal to 1, and K is a process constant.

The 1/f noise is dominant at low frequencies as shown in Figure 4.10. However, beyond the corner frequency (shown as 10 kHz), thermal noise dominates. The effect of 1/f noise on RF circuits can usually be ignored. An exception is in the design of oscillators, where 1/f noise can modulate the oscillator output signal, producing or increasing phase noise. The 1/f noise is also important in direct down-conversion receivers, as the output signal is close to dc. Note also that 1/f noise is much worse for MOS transistors, where it can be significant up to 1 MHz.

4.8 Base Shot Noise Discussion

It is interesting that base shot noise can be related to noise in the resistor r_π by noting that the base shot noise current is in parallel with r_π as shown in Figure 4.11. As shown in (4.25), base shot noise can be related to resistor thermal noise, except that it has a value of $2kTR$ instead of the expected $4kTR$, making use of (4.3), (4.5), (4.7), and (4.22).

$$v_{\text{bn}} = i_{\text{bn}} \times r_\pi = \sqrt{2qI_B} \times r_\pi = \sqrt{2q\frac{I_C}{\beta}} \times r_\pi = \sqrt{2q\frac{I_C}{g_m r_\pi}} \times r_\pi = \sqrt{2q\frac{I_C}{\frac{I_C q}{kT} r_\pi}} \times r_\pi = \sqrt{2kTr_\pi} \tag{4.25}$$

Thus, base shot noise can be related to thermal noise in the resistor r_π (but is off by a factor of 2). This is sometimes expressed by stating that the diffusion resistance

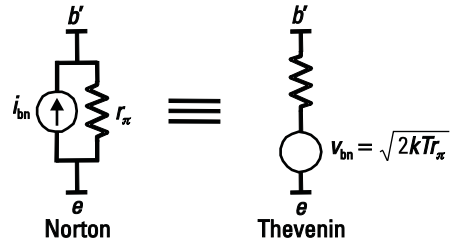

Figure 4.11 Noise model of base shot noise.

is generating noise half thermally. Note that any resistor in thermal equilibrium must generate $\sqrt{4kTR}$ of noise voltage. However, a conducting pn junction is not in thermal equilibrium, and power is added, so it is allowed to break the rules.

4.9 Noise Sources in the Transistor Model

Having discussed the various noise sources in a bipolar transistor, the model for these noise sources can now be added to the transistor model. The noise sources in a bipolar transistor can be shown as in Figure 4.12. These noise sources can also be added to the small-signal model as shown in Figure 4.13.

4.10 Bipolar Transistor Design Considerations

For highest speed, a bias current near peak f_T is suggested. However, it can be noted from Figure 4.8 that the peaks are quite wide. f_T drops by 10% of its peak value

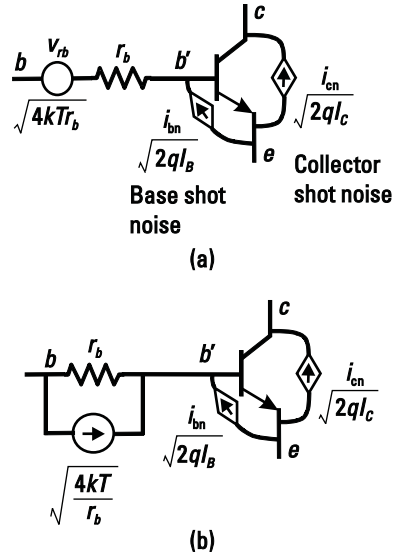

Figure 4.12 Transistor with noise models: (a) with base series noise source and (b) with base parallel noise source.

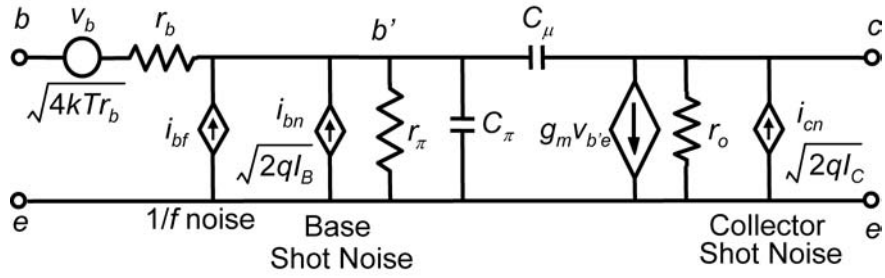

Figure 4.13 Transistor small-signal model with noise.

only when current is reduced to half of the optimum value or when it is increased by 50% over its optimum value. Figure 4.8 also shows that junction capacitance is roughly proportional to transistor size, while base resistance is inversely proportional to transistor size.

Detailed guidelines will be given in the design chapters; however, a few general guidelines are provided as follows:

Pick a lower current to reduce power dissipation with minimal reduction of f_T. As will be further discussed in Chapter 7 for LNA design, the optimal current density for minimum noise figure is also lower than the current required for highest f_T, typically by about a factor of 8, so this is often a good starting point.

Pick a larger transistor size to give lower base resistance. This will have a direct impact on noise. However, on the down side, large size requires large current for optimal f_T. Another negative impact is that junction capacitances increase with larger transistors. The optimal size for best noise performance in an LNA is further discussed in Chapter 7.

Collector shot noise power is proportional to current, but signal power gain is proportional to current squared, so more current can improve noise performance if collector shot noise is dominant. The optimal current for best noise performance in an LNA is further discussed in Chapter 7.

4.11 CMOS Transistors

Bipolar transistors have traditionally been preferred for RF circuits due to the higher values of g_m achievable for a given amount of bias current. However, much more attention is being paid currently to CMOS for the design of radio circuits. Using CMOS allows for the design of single-chip radios, as CMOS is necessary to implement back-end digital or DSP functions. For this purpose, it is possible to use either BiCMOS or straight CMOS. With BiCMOS it is possible to use the bipolar transistors for RF, possibly adding PMOS transistors for power-control functions. However, for economic reasons, or for the need to use a particular CMOS-only process to satisfy the back end requirements, it is now quite common to implement RF circuits in a CMOS only process.

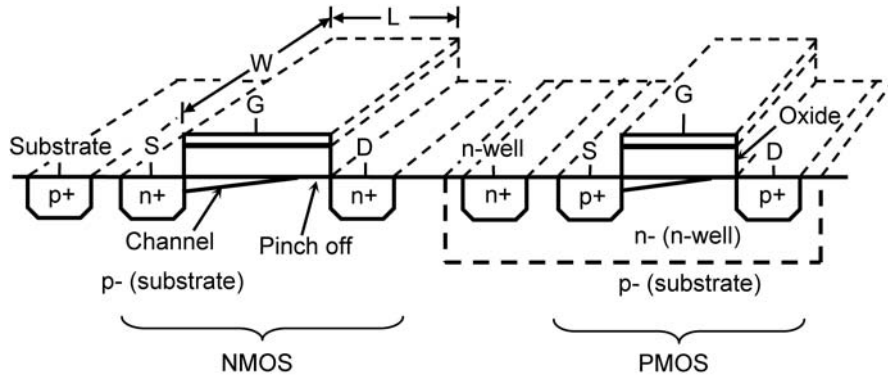

Figure 4.14 CMOS transistors.

We will now provide a brief summary of CMOS transistors. A cross section of basic NMOS and PMOS transistors is shown in Figure 4.14.

The starting material is a low resistivity p substrate. Low resistivity, typically 10 -cm, is used to reduce the likelihood of latch-up in digital circuits. NMOS transistors are formed directly in the p substrate by implanting n+ material to form the source and drain. The region between the source and the drain will later form the channel once the transistor is appropriately biased. On top of the channel is the gate, which is made of polysilicon over a thin layer of oxide. The main reason that polysilicon is typically used instead of metal is that, unlike metal, a desirable threshold voltage can be achieved by appropriately implanting ions into the gate material. Applying a voltage to the gate will control the channel between source and drain as will be described in Section 4.11.1. The substrate contact is made with a region of p+ material implanted into the p substrate.

For the PMOS transistor, an n well is formed in the p substrate as seen in Figure 4.14. In the n well, p+ regions are implanted to form the drain and source. Over the region between the drain and the source, the gate region is formed using poly-silicon over thin oxide. An n+ region in the n-well is used to provide an ohmic contact to the n-well.

Note that all n+ regions in the p substrate form pn diodes. To avoid forward biasing these, the p substrate is kept at the most negative potential. Similarly, p+ regions inside of an n-well form pn diodes. The n-well is kept at the most positive voltage to avoid forward biasing these diodes. Note that in the n-well (these can be separate for every PMOS transistor or for a group of them) it is possible to bias each well separately; for the NMOS transistors, which are in a common substrate, the substrate is connected to the globally most negative voltage. If the substrate is at a lower voltage than the source for an NMOS transistor, it takes a higher gate voltage to form the channel (effectively the threshold voltage is increased) and for a given gate-to-source voltage, the current is reduced. This effect, called the body effect, can be avoided in the PMOS transistor by connecting the n well (the local substrate) to the source. For NMOS transistors, unless the source is grounded, connecting the source to the substrate is not possible as the substrate is common to all NMOS transistors. However, in a process that features a triple well, as depicted in Figure 4.15, source and substrate can be connected,

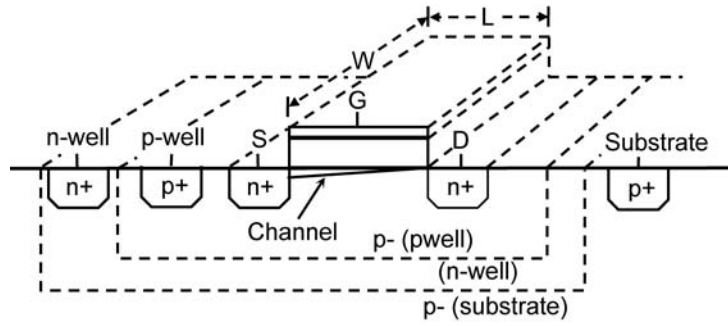

Figure 4.15 Triple well NMOS transistors. Typically, source and p-well (local substrate) are connected, the n-well is connected to the most positive voltage and the (global) p substrate is connected to the most negative voltage.

even if the source is not grounded. Triple well devices are highly desirable for RF applications because of the low transconductance of CMOS devices under even the best circumstances.

4.11.1 NMOS Transistor Operation

The drain characteristic curves for an NMOS transistor are similar to the curves for an npn bipolar transistor and are shown in Figure 4.16. When a positive voltage is applied to the gate of the NMOS device, electrons are attracted towards the gate. With sufficient voltage, an n channel is formed under the oxide, allowing electrons to flow between the drain and the source under the control of the gate voltage v_{GS}. Thus, as gate voltage is increased, more electrons are attracted to the gate surface increasing the electron density, hence the current is increased. For small applied v_{DS}, with constant v_{GS}, a current flows between drain and source due to availability of electrons in the channel. For very low v_{DS}, the relationship is nearly linear as the channel acts like a resistor. For sufficiently large v_{DS}, the channel dimensions at the drain side is reduced to nearly zero as shown in Figure 4.14. The drain voltage at this point is defined as V_{Dsat}. For larger v_{DS}, the drain-source current is saturated

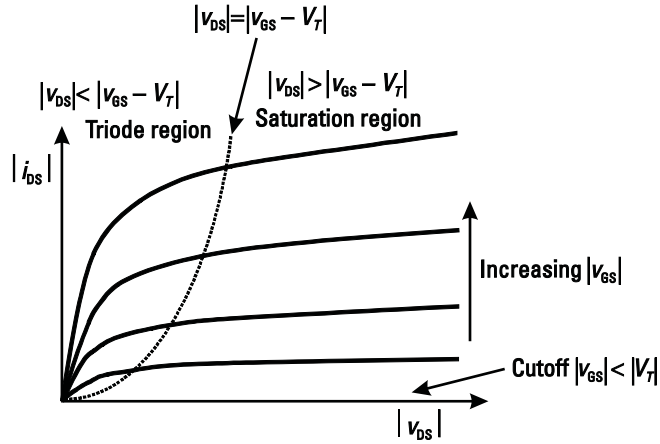

Figure 4.16 CMOS transistor curves. Note values inside the absolute value signs will be positive for NMOS transistors and negative for PMOS transistors.

and remains nearly independent of v_{DS}. This means the output conductance g_{ds} is relatively low, which is advantageous for high gain in amplifiers.

4.11.2 PMOS Transistor Operation

The operation of PMOS is similar to that of NMOS except that negative v_{GS} is applied. This attracts holes, to form a conducting p channel. The characteristic curves for PMOS and NMOS are similar if the absolute value is taken for current and voltage.

4.11.3 CMOS Small-Signal Model

The small-signal model is shown in Figure 4.17. As described in Section 4.11, the current also depends on the voltage from source to substrate—that is, if the source voltage is higher than the substrate voltage, total current is reduced. Such situations would arise in cascode structures, or in a differential pair. The decrease of current due to a nonzero source to substrate voltage, known as the body effect, is modeled by a second current source controlled by v_{sb}. As described in Section 4.11 for transistors in a common substrate, the substrate is biased at the most negative voltage for a p-substrate and at the most positive voltage for an n-substrate. However, if a transistor is placed in its own well, as is standard for PMOS transistors or for NMOS transistors in a triple well, the source can safely be connected to the substrate and the body effect can be eliminated.

As with bipolar transistors, some simplistic equations for calculating model parameters will now be shown. In the saturation region of operation, the current is often described by the simple square law model:

$$i_{DS} = \frac{\mu C_{\text{ox}}}{2} \frac{W}{L} \div (v_{GS} \quad V_T)^2 (1 + \lambda v_{DS}) \tag{4.26}$$

where μ is the carrier mobility, C_{ox} is the gate capacitance per unit area, and V_T is the threshold voltage. λ, the output slope factor, is sometimes described using K, a constant depending on doping concentration, and ϕ_0, the built-in open circuit voltage of silicon [8], as follows:

$$\lambda = \frac{K}{2L\sqrt{V_{DS} \quad (V_{GS} \quad V_T) + \quad _o}} \tag{4.27}$$

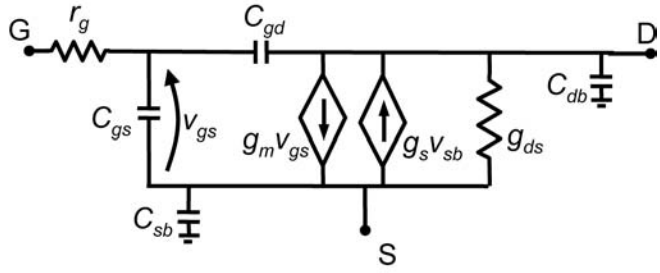

Figure 4.17 Small-signal model for a CMOS transistor. Note the model will be identical for either a PMOS or NMOS device.

λ can be used to calculate the output conductance:

$$g_{ds} = \frac{di_{DS}}{dv_{DS}} = I\lambda \tag{4.28}$$

It is well known that the square law model is too simplistic. More elaborate equations are available, for example including the effects of mobility degradation and velocity saturation effects [4] results in:

$$i_{DS} = \frac{\mu C_{\text{ox}}}{2} \frac{W}{L} \frac{(v_{GS} - V_T)^2}{1 + \alpha(v_{GS} - V_T)^2} (1 + \lambda v_{DS}) \tag{4.29}$$

where α approximately models the combined mobility degradation and velocity saturation effects given by [4]:

$$\alpha = \theta + \frac{\mu_0}{2nv_{\text{sat}}L} \tag{4.30}$$

where θ is the mobility-reduction coefficient and v_{sat} is the saturation velocity. We note that for small values of α or small overdrive voltage ($v_{GS} - V_T$), (4.29) becomes the familiar square law equation as in (4.26).

The transconductance is given by the derivative of the current with respect to the input voltage. For the simple square law equation, this becomes:

$$g_m = \frac{di_{DS}}{dv_{GS}} = \mu C_{\text{ox}} \frac{W}{L} (v_{GS} - v_T)(1 + \lambda v_{DS}) \tag{4.31}$$

This can also be shown to be equal to (note the λ term has been left out):

$$g_m = \sqrt{2\mu C_{\text{ox}} \frac{W}{L} I_{DS}} \tag{4.32}$$

Input capacitance is given by:

$$C_{gs} = \gamma WLC_{\text{ox}} \tag{4.33}$$

The parameter γ is related to fraction of gate not pinched off. A typical value for long-channel γ is 2/3 in saturation and 1 for the case when $V_{DS} = 0$. For a short channel, γ is significantly higher. A typical value is 1, but it has been suggested that values as high as 2 may be appropriate [11].

In the triode region of operation, current is given by:

$$i_{DS} = \mu C_{\text{ox}} \frac{W}{L} \left(v_{GS} - V_T - \frac{v_{DS}}{2} \right) v_{DS} (1 + \lambda v_{DS}) \tag{4.34}$$

In practice, for RF design, short channel devices are used. The above equations for these devices are poor, although performance can still be estimated with

the square law model especially at low overdrive voltages. However, ultimately, it is necessary to use simulators, such as Spectre, ADS, or SPICE, to verify the curves. Recently CMOS models have become much better at predicting transistor performance, although there can still be problems in the simulations of output conductance and phase shift at high frequencies. Other errors can occur due to the improper modeling of parasitics, process variations, coupling between neighboring circuits, or other effects. Thus, measurements are needed to verify any designs, or to refine the models.

4.11.4 f_T and f_{max} for CMOS Transistors

An expression for f_T is the following

$$f_T = \frac{g_m}{2\pi(C_{gs}+C_{gd})} \quad \frac{g_m}{2\pi C_{gs}} = \frac{\mu C_{ox} \quad \frac{W}{L} \div (v_{GS} \quad v_T)}{2\pi\gamma W L C_{OX}} = \frac{\mu(v_{GS} \quad v_T)}{2\pi\gamma L^2} \tag{4.35}$$

where γ 2/3 in saturation for a long-channel device, and higher for short channel devices.

An expression for f_{max}, similar to the expression for a bipolar transistor derived earlier, is [11]:

$$f_{max} = \sqrt{\frac{f_T}{8\pi r_g C_{gd}}} \tag{4.36}$$

Given the expression for f_{max}, it is also possible to estimate for the maximum achievable power gain for an amplifier built with this transistor as follows [11]

$$G_{max} = \frac{f_T}{8\pi r_g C_{gd} f^2} \tag{4.37}$$

4.11.5 CMOS Small-Signal Model Including Noise

Similar to the bipolar transistor, the thermal channel noise in a MOSFET can be modeled by placing a noise current source, i_{nd} in parallel with the output as shown in Figure 4.18.

$$i_{nd}^2 = 4kT\gamma g_m \tag{4.38}$$

As before, the parameter γ is about 2/3 for a long channel device in saturation, and larger (e.g., 1 to 2) for short channel devices.

Gate resistance noise can be modeled with voltage source v_{ng}.

$$v_{ng}^2 = 4kTr_g \tag{4.39}$$

In addition, CMOS transistors experience noise due to the distributed nature of the transistor. Such nonquasistatic effects not only cause a decrease in f_{max} but

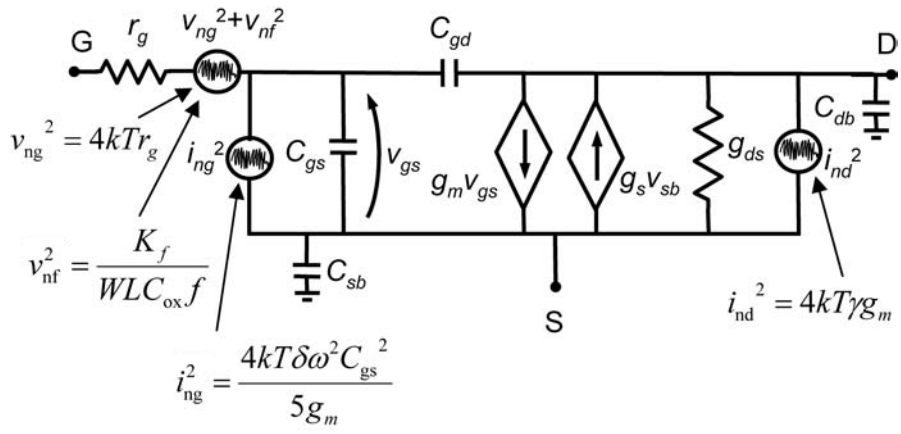

Figure 4.18 CMOS small-signal model with noise.

also a significant noise contribution, especially as frequencies increase past about a tenth of f_T. This noise, sometimes called induced gate noise, can be modeled as an additional current i_{ng} in Figure 4.18 and is described by

$$i_{ng}^2 = \frac{4kT\delta\omega^2 C_{gs}^2}{5g_m} \tag{4.40}$$

A typical value for parameter δ is about twice the value of γ or about 4/3 for a long channel and from 2 to 4 for a short channel [11]. Finally, 1/*f* noise can be expressed with

$$v_{nf}^2 = \frac{K_f}{WLC_{ox}f} \tag{4.41}$$

where K_f is a process constant.

For analog design, it is often assumed that the input impedance Z_{in} is infinity, in which case the input noise current would be zero. However, at RF, both input referred noise current and noise voltage are required to account for the actual RF input impedance, which is neither zero nor infinity. Another point to note is that input noise voltage and noise current, as shown in Figure 4.19, are correlated sources, since they both have the same origin. This is similar to the case in bipolar transistors with collector shot noise referred to the base.

We note that gate resistance in a CMOS transistor is an important source of noise as well as an important factor in determining f_{max}. Gate resistance can be calculated from the dimensions of the gate and the gate sheet resistance ρ_s by:

$$R_{GATE} = \frac{1}{3}\rho_s \frac{W}{L} \tag{4.42}$$

for a gate with a contact on one side. Typical values for ρ_s can be from a few ohms per square to more than 10 ohms per square. We note that the gate poly by itself

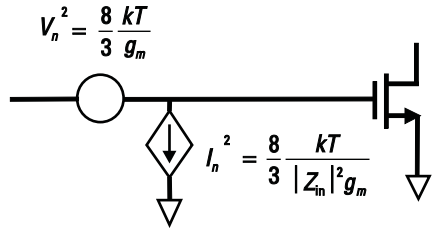

Figure 4.19 Gate-referred noise in NMOS transistor. Only noise due to I_{nd} is shown.

would have a resistance of $\rho_s W/L$. The factor of 1/3 in (4.42) comes from the distributed resistance in the gate and the fact that the transistor current is flowing under all regions of the gate. The series resistance varies from 0 near the contact to $\rho_s W/L$ for the far end of the gate, with an effective value given by (4.42). If the gate is contacted on both sides, the effective resistance drops by a further factor of 4 such that:

$$R_{\text{GATE}} = \frac{1}{12}\rho_s \frac{W}{L} \tag{4.43}$$

Up until recently, some of these noise sources were not included in simulator models and as a result simulations tended to predict optimistic noise performance of CMOS transistors. However, with newer models such as BSIM4, gate noise is better modeled and realistic results can be obtained. However, the designer should be aware that there are switches in some design kits to turn gate resistance or particular noise models on or off, and this can have a large impact on results or even on design techniques. As an example, if gate resistance is turned off, the number of gate fingers used will have much less impact on circuit noise; hence, the designer may be tempted to use fewer fingers. With noise properly modeled, typically individual gate widths in modern processes should not exceed a few microns. Instead, multiple fingers are used to achieve minimum noise with the desired overall transistor size. Finally, it should be noted that noise modeling in CMOS transistors is still an area of active research.

All of the above numbers will scale with transistor size and suggests an approach for designing by selecting the current density to result in desired maximum operating frequency and lowest noise. It has been shown by [12] that measured optimum f_T, $f_{\max}$, and $NF_{\min}$ typically occur at a bias current in mA numerically equal to about 0.2 times the gate width in microns, for a number of different CMOS technologies. This was demonstrated across different process nodes, from different vendors, at different channel lengths, and for different circuit topologies. This is in rough agreement with our simulated results, however, it should be noted that unless the process comes with well characterized, full RF circuit models, simulations may not be accurate. Our simulation results showed differences between cascode structures and common source transistors, in large part likely due to voltage limiting.

Example 4.2: Transistor Models

For a 0.13-μm CMOS process, assume C_{ox} is 10 fF/(μm)^2, electron mobility is 500 cm^2/V-sec, and gate poly has a sheet resistance of 7.6 /square. Assume the

effective minimum gate length is 0.10 μm. For a minimum length transistor with a gate width of 20 μm, compare simple square law predictions to simulation results for current, transconductance, and noise versus bias current. Also show results for f_T, $f_{\max}$, and transistor noise figure NF.

Solution:
Square law predictions for current, ignoring output conductance, result in

$$i_{DS} = \frac{\mu C_{ox}}{2} \frac{W}{L} \div (v_{GS} \quad V_T)^2 = \frac{500 \text{ cm}^2/V \text{ sec} \quad 1 \ \mu\text{F/cm}^2}{2} \frac{20}{0.1} (v_{GS} \quad V_T)^2$$

$$= 50 \text{ mA/V}^2 (v_{GS} \quad V_T)^2$$

Figure 4.20 shows the simulated current I_{DS} versus v_{GS} for the same transistor in a typical 130 nm CMOS process. It turns out the best fitting square law curve has a constant of 29 mA/V^2 rather than the predicted 50 mA/V^2. This is likely due to reduction in mobility due to the electric field and velocity saturation but indicates the importance of doing some simple transistor characterization rather than relying on simple calculations.

The derivative of this curve has been taken in Figure 4.20 to show the transconductance g_m versus v_{GS}. It should be noted that the values for g_m are in agreement with ac values all the way up to 20 GHz, although at 20 GHz, the output current has some significant phase shift with respect to the input voltage. It should be noted that square law is a good model to predict current up to a v_{GS} of about 0.7V and up to about 0.5V for predicting transconductance.

A prediction for f_T was given in (4.35). Using the value for g_m at high overdrive voltages where f_T tends to be highest results in

$$f_T \quad \frac{g_m}{2\pi C_{gs}} = \frac{15 \text{ mA/V}}{2\pi \quad 20\mu \quad 0.1\mu \quad 10 \text{ fF}/\mu\text{m}^2} \quad 120 \text{ GHz}$$

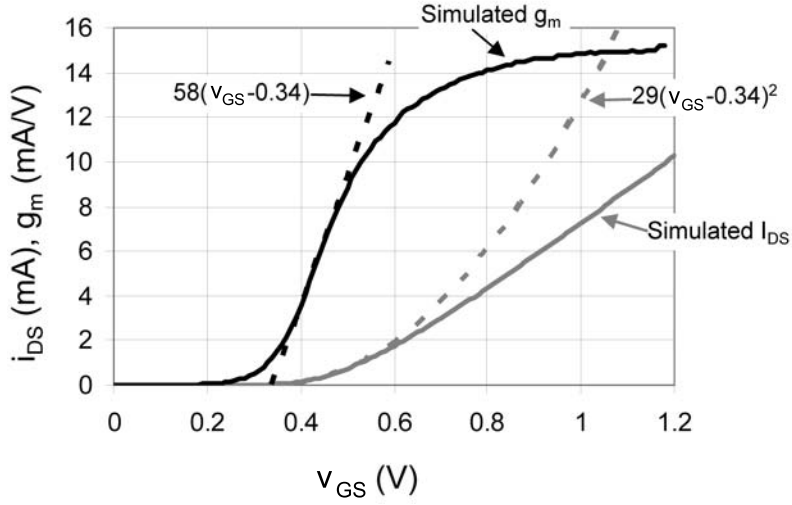

Figure 4.20 Simulated current and transconductance compared to square law equations.

In this calculation, the factor of γ has been left out. Also, the omission of C_{gd} can be quite important as it could be as high as 30% of C_{gs} and this would reduce f_T by the same amount. A prediction for $f_{\max}$ is given by (4.36). To determine numbers we need to estimate gate resistance. With the given sizes and using (4.43), we determine

$$R_{\text{GATE}} = \frac{1}{12}\rho_s \frac{W}{L} = \frac{1}{12} 7.6 \frac{20}{square\ 0.1} = 127$$

Estimating C_{gd} as 7 fF and using the above value for f_T, $f_{\max}$ can be estimated as

$$f_{\max} = \sqrt{\frac{f_T}{8\pi r_g C_{gd}}} = \sqrt{\frac{120\ \text{GHz}}{8\pi \quad 127 \quad 7\text{fF}}} = 73\ \text{GHz}$$

The large value of gate resistance will limit $f_{\max}$ and will also have a huge impact on noise. Recall that a 50 resistor would add 3 dB to the noise, so clearly gate resistance must be reduced. If instead of using 1 finger with a width of 20 μm the gate is split into 20 fingers in parallel each with a width of 1 μm, then each finger resistance is reduced by a factor of 20. Furthermore, by placing all 20 in parallel, there is a further reduction of resistance by a factor of 20, thus resistance is reduced by a factor of 400 down to 0.316 . This will be less than the internal resistance and thus will no longer limit the $f_{\max}$ and the noise; it also cannot be used to calculate $f_{\max}$.

Figure 4.21 shows f_T and $f_{\max}$ as a function of bias current for a 20-μm transistor with 20 fingers. The circuit was driven from an ac-coupled 50 input and output. The load was a large inductor in parallel with a noise-free 5-k resistor and bias was supplied with a current mirror isolated from the circuit with a 10-k resistor. The curve for f_T, was obtained as the unity gain frequency of H_{21} versus frequency, while $f_{\max}$ was obtained from the unity gain frequency of $G_{\max}$ versus frequency. It should be noted that the value for f_T, shown as higher than 100 GHz, is quite optimistic as even minimal connections to the various terminals will add a substantial amount of capacitance. It can also be observed that the peak for $f_{\max}$

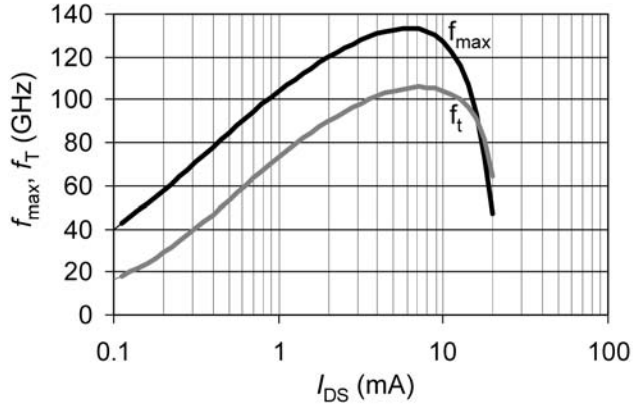

Figure 4.21 Simulated f_T and $f_{\max}$ for common-source transistors.

and f_T occur at about 7 and 8 mA, respectively; from Figure 4.20 this requires a v_{GS} approaching that of the power supply voltage, and thus this will not be a useful operating point when more than one transistor is connected in series between the power supply rails. This point will be elaborated on in Chapter 8 when amplifiers are discussed.

Noise factor and noise figure can be determined for a transistor by comparing the total noise sources to the noise from the source as discussed in Chapter 2. The output noise current squared due to the source, gate resistance, drain current noise, and load resistor is given by

$$i_{no}{}^2 = (4kTR_S + 4kTR_g)g_m{}^2 + 4kTg_m\gamma + \frac{4kT}{R_L}$$

If this is referred back to the input by dividing by $g_m{}^2$ and normalizing to the source noise, the resulting noise factor is

$$F = 1 + \frac{R_g}{R_S} + \frac{\gamma}{g_m R_S} + \frac{1}{g_m{}^2 R_S R_L}$$

It should be noted that in this example, noise from the load resistor was not included, but is shown in the above equations for completeness. Figure 4.22 shows the predicted and simulated noise figure considering only source noise, gate resistance noise, and drain channel noise for both 1 finger and for 20 fingers showing the significant difference. For the predicted noise, the dashed line was determined using a constant γ of 1 and the square law equation for g_m. As g_m keeps growing with increasing current, the noise decreases; however, it is noted that the simulated noise comes to a minimum at about 1 mA then rises again. The reason is that the actual g_m saturates and no longer increases. As well, γ increases from a value of about 0.5 in the subthreshold region to a value of 2 or larger for large overdrive voltages. The dotted lines in Figure 4.22 show the calculated noise figure using the simulated g_m and γ that increases linearly with V_{GS} from 0.5 at the threshold to 2.5 at V_{DD}. Since

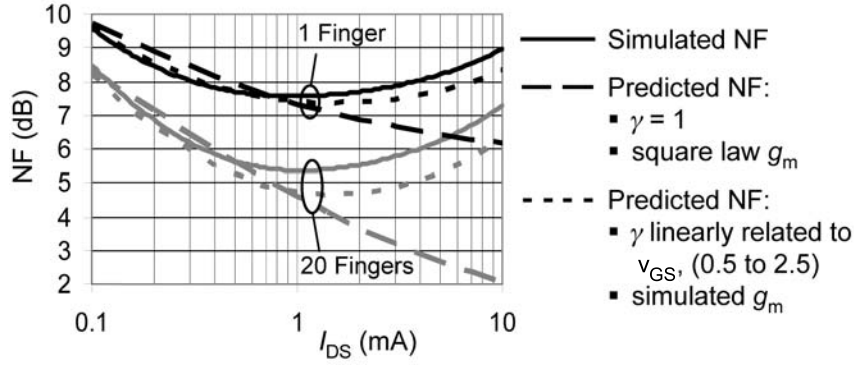

Figure 4.22 Predicted and simulated noise figure for common-source transistor with 1 finger and 20 fingers and comparing constant and square law g_m with variable and simulated g_m.

g_m stops increasing, the noise figure no longer falls and because $\gamma$is increasing, noise figure actually starts to rise at high overdrive voltages and the shape of the curve approximately matches that of the simulation. By performing a noise summary it was determined that the remaining discrepancy is due mainly to $1/f$ noise and noise due to source and substrate resistances.

In Chapter 7, it will be shown how to use scaling and matching to achieve considerably lower noise than shown in this example.

4.12 Practical Considerations in Transistor Layout

4.12.1 Typical Transistors

A typical layout for a transistor is shown in Figure 4.23.

Typically standard transistors have been optimized for the highest-frequency response, fast switching, and compact size for digital circuits. However, most processes also have low and high threshold devices. Low threshold devices may have higher leakage when off, but allow for operation at very low supply voltages. Higher threshold transistors are built with thicker gate oxide; hence, will have higher breakdown voltage, especially useful in the design of power amplifiers. However, with thicker oxide, typically minimum gate lengths are not allowed, hence the f_T is reduced so higher breakdown voltages come at the expense of frequency response. Another option is the triple well transistor. As discussed previously, triple well transistors will minimize the body effect and hence it is easier to build followers with better drive capability. In terms of matching, proximity is important. Transistors in their own separate wells will be further apart hence matching can be worse.

4.12.2 Symmetry

In differential circuits, symmetry is very important to ensure that unwanted coupling into both sides of the circuit is equal. This means the coupled signal will be

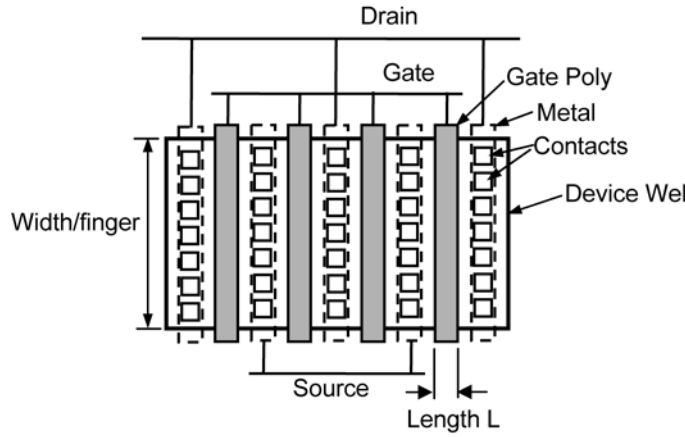

Figure 4.23 Layout for CMOS transistor with four fingers.

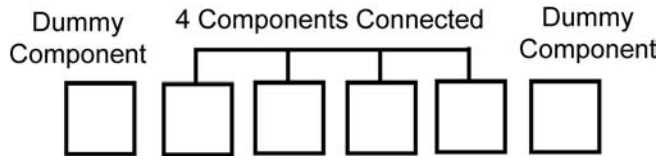

Figure 4.24 Dummy components to reduce edge effects.

a common-mode signal and can be rejected. Symmetrical circuits also prevent unequal signal delays again ensuring that signals are properly lined up.

4.12.3 Matching

To achieve good matching it is important that components being matched are close together and are oriented in the same direction. When multiple components (for example, resistors, capacitors, or transistor fingers) need to be matched, additional dummy components are sometimes added at the edges, as shown in Figure 4.24. Dummy components are not connected to the circuit, but are there to reduce edge effects.

A common way to obtain better matching between several components is to interleave subparts of these components. If these components are transistor fingers, and if the source is common to both the transistors, they can be placed in the same device well as shown in Figure 4.25.

Another technique is to place components such as transistors in a common centroid configuration, in which each component is a parallel combination of subcomponents, which are arranged in a pattern on the die so that process variations in the x or y direction will affect subcomponents of each transistor. As an example, two transistors M1 and M2 each of size W/L can be made up of two subtransistors each of size $0.5\ W/L$ as shown in Figure 4.26. In this way process variation in the x or y directions will affect each transistor in the same way and matching will be maintained. However, the common centroid approach results in the need for more interconnect resulting in more parasitic capacitance, and as a result frequency response may be poorer.

As another example of a tradeoff, triple-well transistors can provide more drive at low voltages, but typically may have worse matching because of the typical requirement of more spacing between such transistors compared to regular transistors.

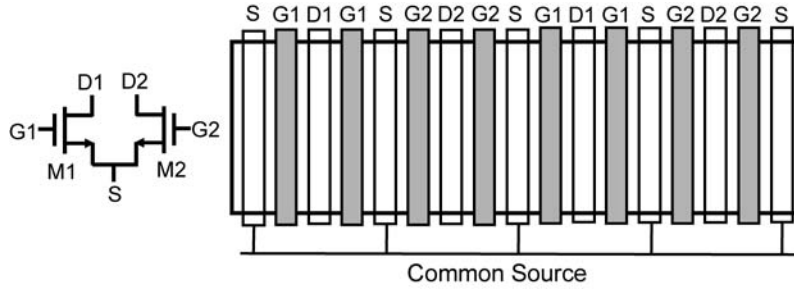

Figure 4.25 Interleaved transistors having a common source. Each transistor has four fingers.

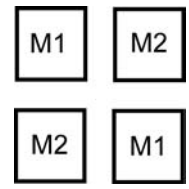

Figure 4.26 Common-centroid arrangement of M1 and M2 reducing the effect of processing variations in either *x* or *y* directions.

4.12.4 ESD Protection and Antenna Rules

Transistor performance can be affected by the need to have protection against electrostatic discharge or ESD. ESD can be the result of external factors such as static charge. It can also be the result of processing steps, such as plasma etching. When a significant amount of metal is attached to a gate, processing of that metal during fabrication can result in charge buildup on the gate and ultimately the destruction of the gate even before the circuit is delivered back to the designer. Such failures can be avoided by limiting the total area of metal relative to the gate area, or by adding protection circuitry such as diodes to the substrate. Rules to avoid such breakdown of gate oxides are called antenna rules.

References

[1] Taur, Y., and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge, U.K.: Cambridge University Press, 1998.

[2] Plummer, J. D., P. B. Griffin, and M. D. Deal, *Silicon VLSI Technology: Fundamentals, Practice, and Modeling*, Upper Saddle River, NJ: Prentice-Hall, 2000.

[3] Sedra, A. S., and K. C. Smith, *Microelectronic Circuits*, 5th ed., New York: Oxford University Press, 2004.

[4] Terrovitis, M. T., and R. G. Meyer, "Intermodulation Distortion in Current-Commutating CMOS Mixers," *IEEE J. of Solid-State Circuits*, Vol. 35, No. 10, October 2000, pp. 1461–1473.

[5] Roulston, D. J., *Bipolar Semiconductor Devices*, New York: McGraw-Hill, 1990.

[6] Streetman, B. G., *Solid-State Electronic Devices*, 3rd ed., Englewood Cliffs, NJ: Prentice-Hall, 1990.

[7] Muller, R. S., and T. I. Kamins, *Device Electronics for Integrated Circuits*, New York: John Wiley & Sons, 1986

[8] Sze, S. M., *High Speed Semiconductor Devices*, New York: John Wiley & Sons, 1990.

[9] Sze, S. M., *Modern Semiconductor Device Physics*, New York: John Wiley & Sons, 1997.

[10] Cooke, H., "Microwave Transistors Theory and Design," *Proc. IEEE*, Vol. 59, August 1971, pp. 1163–1181.

[11] Lee, T. H., *The Design of CMOS Radio-Frequency Integrated Circuits*, 2nd ed., Cambridge, U.K.: Cambridge University Press, 2004.

[12] Dickson, T. O., et al., "The Invariance of Characteristic Current Densities in Nanoscale MOSFETs and Its Impact on Algorithmic Design Methodologies and Design Porting of Si(Ge) (Bi)CMOS High-Speed Building Blocks," *J. Solid-State Circuits*, Vol. 41, No. 8, August 2006, pp. 1830–1845.

CHAPTER 5

Impedance Matching

5.1 Introduction

In RF circuits, we very seldom start with the impedance that we would like. Therefore, we need to develop techniques for transforming an arbitrary impedance into the impedance of choice. For example, consider the RF system shown in Figure 5.1. Here the source and load are 50 (a very popular impedance), as are the transmission lines leading up to the IC. For optimum power transfer, prevention of ringing and radiation, and good noise behavior, for example, we need the circuit input and output impedances matched to the system. In general, some matching circuit must usually be added to the circuit, as shown in Figure 5.2.

Typically, reactive matching circuits are used because they do not add noise to the circuit. However, using reactive matching components means that the circuit will only be matched over a range of frequencies and not at others. If a broadband match is required, then other techniques may need to be used. An example of matching a transistor amplifier with a capacitive input is shown in Figure 5.3. The series inductance adds an impedance of $j\omega L$ to cancel the input capacitive impedance. Note that in general, when an impedance is complex ($R + jX$), to match it, the impedance must be driven from the complex conjugate ($R - jX$).

A more general matching circuit is required if the real part is not 50 . For example, if the real part of Z_{in} is less than 50 , then the circuit can be matched using the circuit in Figure 5.4 and described in Example 5.1.

Example 5.1: Matching Using Algebra Techniques

A possible impedance matching network is shown in Figure 5.4. Use the matching network to match the transistor input impedance $Z_{in} = 40 - j30$ to $Z_o = 50$. Perform the matching at 2 GHz.

Solution:

We can solve for Z_2 and Y_3, where for convenience, we have chosen impedance for series components and admittance for parallel components. An expression for Z_2 is:

$$Z_2 = Z_{in} + j\omega L$$

where $Z_{in} = R_{in} - jX_{in}$. Solving for Y_3 and equating it to the reference admittance Y_o:

$$Y_3 = Y_2 + j\omega C = \frac{1}{Z_o} = Y_o$$

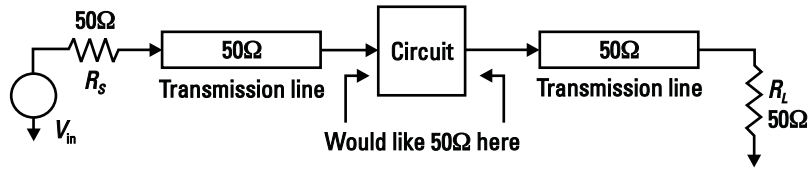

Figure 5.1 Circuit embedded in a 50 system.

Using the above two equations, Y_2 is eliminated, leaving only L and C as unknowns:

$$\frac{1}{Z_{\text{in}} + j\omega L} = Y_o \quad j\omega C$$

Solving the real and imaginary parts of this equation, values for C and L can be found. With some manipulation:

$$\frac{R_{\text{in}} \quad j(\omega L \quad X_{\text{in}})}{R_{\text{in}}{}^2 + (\omega L \quad X_{\text{in}})^2} = Y_o \quad j\omega C$$

The real part of this equation gives

$$\omega L = X_{\text{in}} + \frac{\sqrt{R_{\text{in}} \quad Y_o R_{\text{in}}^2}}{Y_o} = 30 + \frac{\sqrt{40 \quad (0.02)(40)^2}}{0.02} = 50$$

Now using the imaginary half part of the equation:

$$\omega C = \frac{\omega L \quad X_{\text{in}}}{R_{\text{in}}^2 + (\omega L \quad X_{\text{in}})^2} = \frac{50 \quad 30}{40^2 + (50 \quad 30)^2} = 0.01$$

At 2 GHz it is straightforward to determine that L is equal to 3.98 nH and C is equal to 796 fF. We note that the impedance is matched exactly only at 2 GHz. We also note that this matching network cannot be used to transform all impedances to 50 . Other matching circuits will be discussed later.

Although the preceding analysis is very useful for entertaining undergraduates during final exams, in practice there is a more general method for determining a matching network and finding the values. However, first we must review the Smith chart.

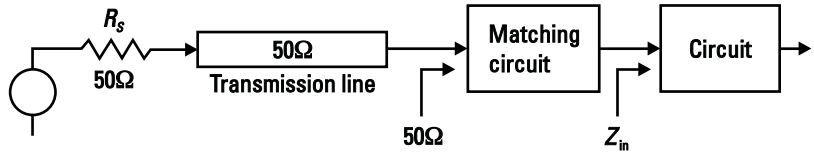

Figure 5.2 Circuit embedded in a 50 system with matching circuit.

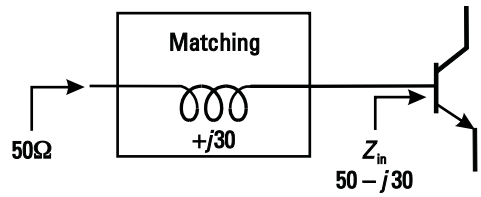

Figure 5.3 Example of a very simple matching network.

5.2 Review of the Smith Chart

The reflection coefficient is a very common figure of merit used to determine how well matched two impedances are. It is related to the ratio of power transmitted to power reflected from the load. A plot of the reflection coefficient is the basis of the Smith chart, which is a very useful way to plot the impedances graphically. The reflection coefficient can be defined in terms of the load impedance Z_L and the characteristic impedance of the system Z_o as follows:

$$= \frac{Z_L \quad Z_o}{Z_L + Z_o} = \frac{z \quad 1}{z+1} \tag{5.1}$$

where $z = Z_L/Z_o$ is the normalized impedance. Alternatively, given , one can find Z_L or z as follows:

$$Z_L = Z_o \frac{1+}{1} \tag{5.2}$$

or

$$z = \frac{1+}{1} \tag{5.3}$$

For any impedance with a positive real part, it can be shown that:

$$0 \quad | \; | \quad 1 \tag{5.4}$$

The reflection coefficient can be plotted on the x-y plane and its value for any impedance will always fall somewhere in the unit circle. Note that for the case where $Z_L = Z_o$, = 0. This means that the center of the Smith chart is the point

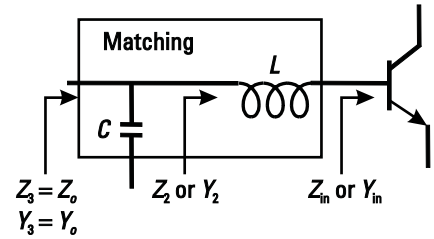

Figure 5.4 A possible impedance matching network.

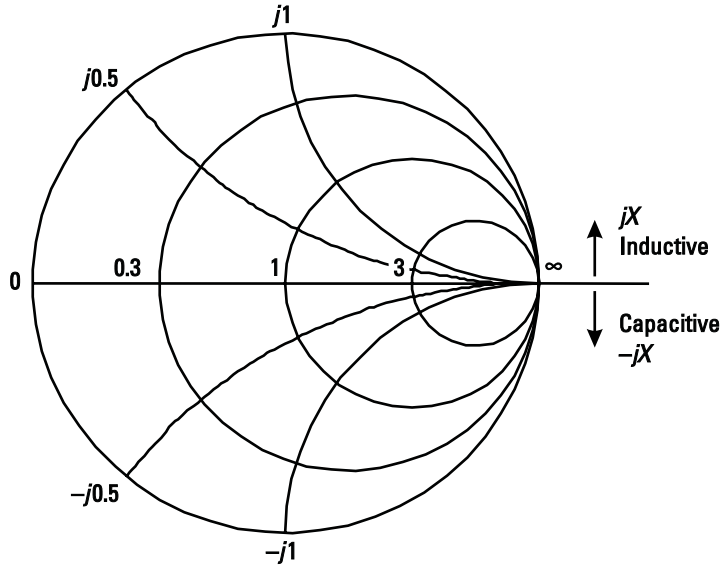

Figure 5.5 A Smith chart.

where the load is equal to the characteristic impedance (in other words perfect matching). Real impedances lie on the real axis from 0 at = 1 to at = +1. Purely reactive impedances lie on the unit circle. Thus, impedances can be directly shown, normalized to Z_o. Such a plot is called a Smith chart and is shown in Figure 5.5. Note that the circular lines on the plot correspond to contours of constant resistance while the arcing lines correspond to lines of constant reactance. Thus, it is easy to graph any impedance quickly. Table 5.1 shows some impedances that can be used to map out some important points on the Smith chart (it assumes that $Z_o = 50$).

Just as contours of constant resistance and reactance were plotted on the Smith chart, it is also possible to plot contours of constant conductance and susceptance. Such a chart can also be obtained by rotating the Z chart (impedance Smith chart) by 180°. An admittance Smith chart, or Y chart, is shown in Figure 5.6. This set of admittance curves can be overlaid with the impedance curves to form a YZ Smith chart as shown in Figure 5.7. Then for any impedance Z, the location on the chart can be found, and Y can be read directly or plotted. This will be shown next to be useful in matching, where series or parallel components can be added to move the impedance to the center, or to any desired point.

Table 5.1 Mapping Impedances to Points on the Smith Chart

Z_L	
50	0
0	1
	1
100	0.333
25	0.333
$j50$	1 90°
jX	1 2 tan $^{1}(X/50)$
$50 - j141.46$	0.8166 35.26°

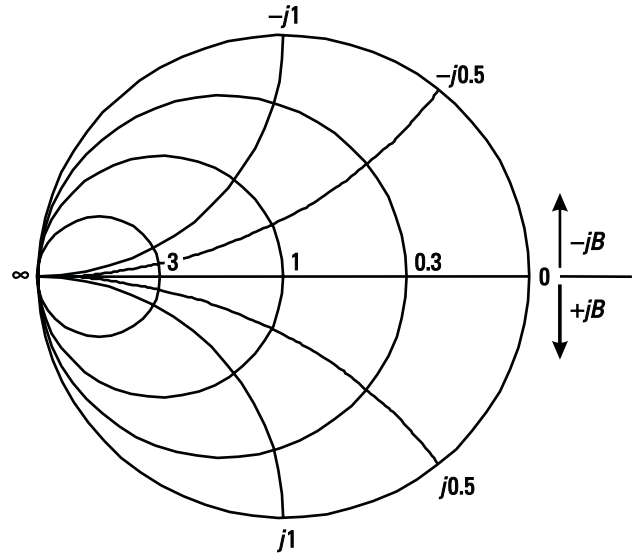

Figure 5.6 An admittance Smith chart or a Y Smith chart.

In addition to the reflection coefficient there are a number of other related values that can be calculated. The voltage standing wave ratio (VSWR) is defined as:

$$VSWR = \frac{1+\left|\ \right|}{1\ \left|\ \right|} = \frac{|v_x|_{\max}}{|v_x|_{\min}} \tag{5.5}$$

where $|v_x|_{\max}$ and $|v_x|_{\min}$ are the maximum and minimum magnitudes of the standing wave on a transmission line.

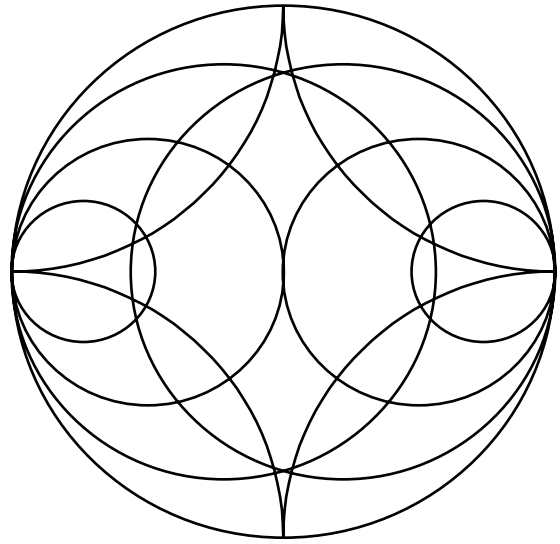

Figure 5.7 A YZ Smith chart.

Table 5.2 Using Lumped Components to Match Circuits

Component Added	*Effect*		*Description of Effect*
Series inductor	z	$z + j\omega L$	Move clockwise along a resistance circle.
Series capacitor	z	$z \quad j/\omega C$	Smaller capacitance increases impedance ($j/\omega C$) to move counterclockwise along a resistance circle.
Parallel inductor	y	$y \quad j/\omega L$	Smaller inductance increases admittance ($j/\omega L$) to move counterclockwise along a conductance circle.
Parallel capacitor	y	$y + j\omega C$	Move clockwise along a conductance circle.

Another value that is often calculated related to the reflection coefficient is called the return loss (RL) and it is given by:

$$RL = 20 \log \frac{VSWR + 1}{VSWR \quad 1} \div = 20 \log \frac{1}{\left| \right|} \div \tag{5.6}$$

Note if the system is perfectly matched, the RL is minus infinity and the VSWR would be one.

5.3 Impedance Matching

The input impedance of a circuit can be any value. In order to have the best power transfer into the circuit, it is necessary to match this impedance to the impedance of

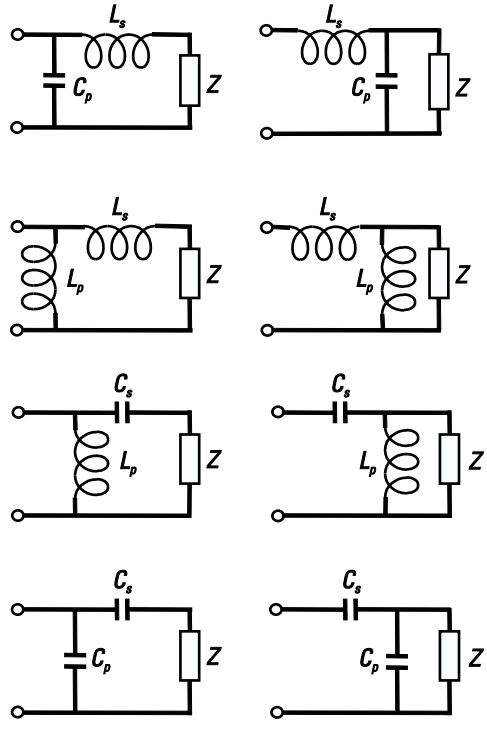

Figure 5.8 The eight possible impedance matching networks with two reactive components.

the source driving the circuit. The output impedance must be similarly matched. It is very common to use reactive components to achieve this impedance transformation because they do not absorb any power or add noise. Thus, series or parallel inductance or capacitance can be added to the circuit to provide an impedance transformation. Series components will move the impedance along a constant resistance circle on the Smith chart. Parallel components will move the admittance along a constant conductance circle. Table 4.2 summarizes the effect of each component.

With the proper choice of two reactive components, any impedance can be moved to a desired point on the Smith chart. There are eight possible two-component matching networks, also known as *ell* networks, as shown in Figure 5.8. Each will have a region in which a match is possible, and a region in which a match is not possible.

In any particular region on the Smith chart, several matching circuits will work and others will not. This is illustrated in Figure 5.9, which shows what matching networks will work in what regions. Since more than one matching network will work in any given region, how does one choose? There are a number of popular reasons for choosing one over another.

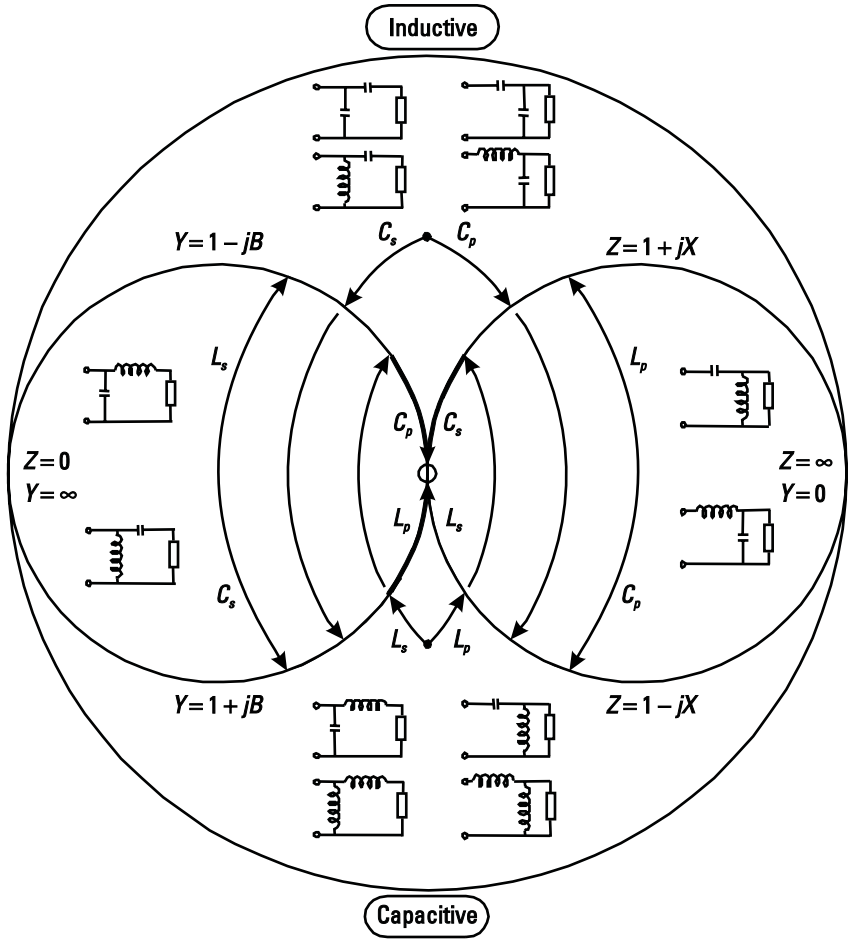

Figure 5.9 Diagram showing which ell matching networks will work in what regions.

1. Sometimes matching components can be used as dc blocks (capacitors) or to provide bias currents (inductors).
2. Some circuits may result in more reasonable component values.
3. Personal preference. Not to be underestimated, sometimes when all paths look equal you just have to shoot from the hip and pick one.
4. Stability. Since transistor gain is higher at lower frequencies, there may be a low frequency stability problem. In such a case, sometimes a high-pass network (series capacitor, parallel inductor) at the input may be more stable.
5. Harmonic filtering can be done with a lowpass matching network (series L, parallel C). This may be important for power amplifiers, for example.

Example 5.2: General Matching Example

Match $Z = 150 \quad 50j$ to $50 \quad$ using the techniques just developed.

Solution:
We first normalize the impedance to $50 \quad$. Thus the impedance that we want to match is $3 \quad 1j$. We plot this on the Smith chart as point A as shown in Figure 5.10. Now we can see from Figure 5.9 that in this region we have two possible matching networks. We choose arbitrarily to use a parallel capacitor and then a

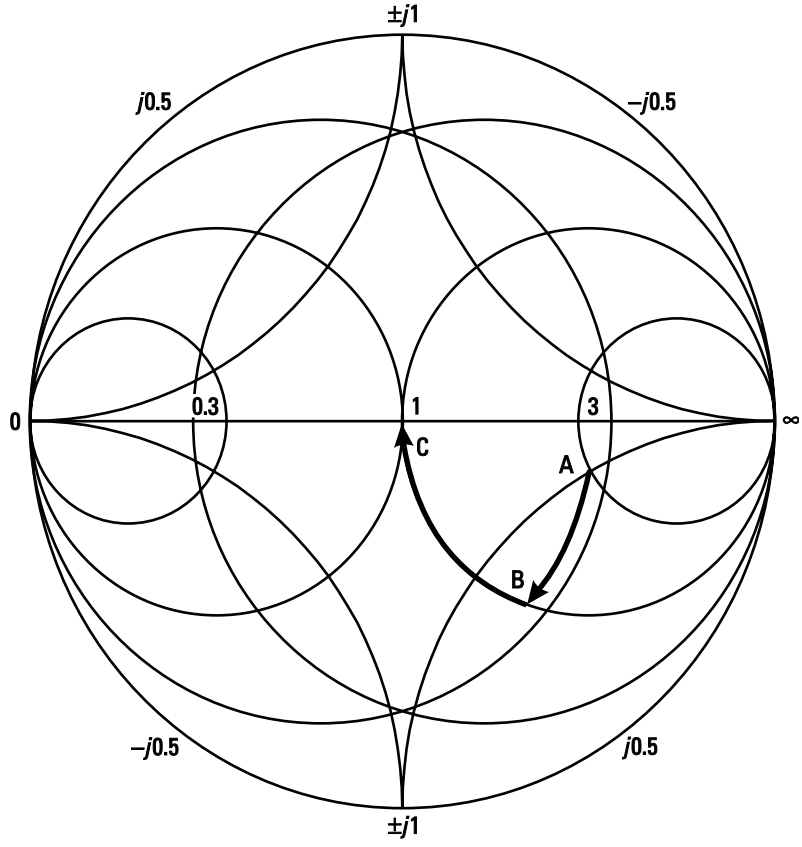

Figure 5.10 Illustration of matching process.

series inductor. Adding a parallel capacitor moves the impedance around a constant admittance circle to point *B*, which places the impedance on the 50 resistance circle. Once on the unit circle, a series inductance moves the impedance along a constant resistance circle and moves the impedance to the center at point *C*. The values can be found by noting that point *A* is at $Y_A = 1/Z_A = 0.3 + 0.1\,j$, and *B* is at $Y_B = 0.3 + 0.458j$; therefore we need a capacitor admittance of $0.358j$. Since $Z_B = 1/Y_B =$ 1 $1.528j$ an inductor reactance of $1.528j$ is needed to bring it to the center.

Example 5.3: Illustration of Different Matching Networks

Match a 200 load to a 50 source at 1 GHz with both a lowpass and a highpass matching network, illustrating the filtering properties of the matching network.

Solution:
Using techniques as above, two matching circuits—as shown in Figure 5.11—are designed. The frequency response can be determined with results as shown in Figure 5.12. It would seem from this diagram that for the lowpass matching network, the signal can be transferred from dc to the –3-dB corner at about 1.53 GHz. However, as seen in the plot of the input impedance in Figure 5.13, the impedance is only matched in a finite band around the center frequency. For the lowpass network, the impedance is within 25 of 50 from 0.78 to 1.57 GHz. It can be noted that if the difference between the source and the load is higher, the bandwidth of the matching circuit will be narrower.

For optimal power transfer and minimal noise, impedance should be controlled (although the required impedance for optimal power transfer may not be the value for minimal noise). Also, sources, loads, and connecting cables or transmission lines will be at some specified impedance typically 50 .

Example 5.4: The Effect of Matching on Noise

Study the impact of matching on base shot noise. Use the 15 transistor defined in Table 4.1, which is operated at 2 mA at 1 GHz. Assume C_π remains at 700 fF.

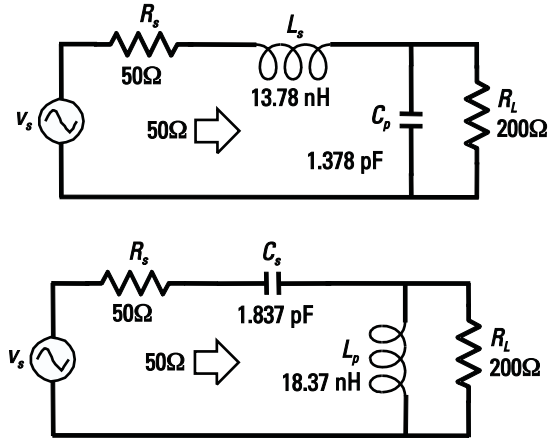

Figure 5.11 Lowpass and highpass matching network for Example 5.4.

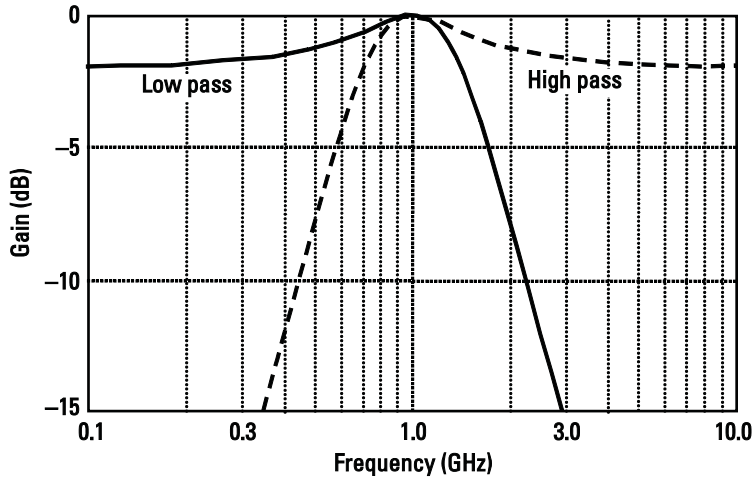

Figure 5.12 Frequency response for lowpass and highpass matching networks.

Solution:
The small signal model and calculated matching impedances are shown in Figure 5.14. The transistor has an input impedance of 1,250 in parallel with 700 fF, which at 1 GHz is equal to $Z_\pi = 40 \quad j220$. Using this and the base resistance of 5 , the impedance seen by the base shot noise source can be determined. Without matching, the input is driven by 50 , so the base shot noise sees 55 in parallel with 40 $j220$, which is equal to about 50 $j11.6$. With matching, the base shot noise sees $50 + j220$ in parallel with 40 $j220$ for a net impedance of 560 $j24$. Thus with matching, the base shot noise current sees an impedance whose magnitude is about 10 times higher, thus the impact of the base shot noise is significantly worse with impedance matching.

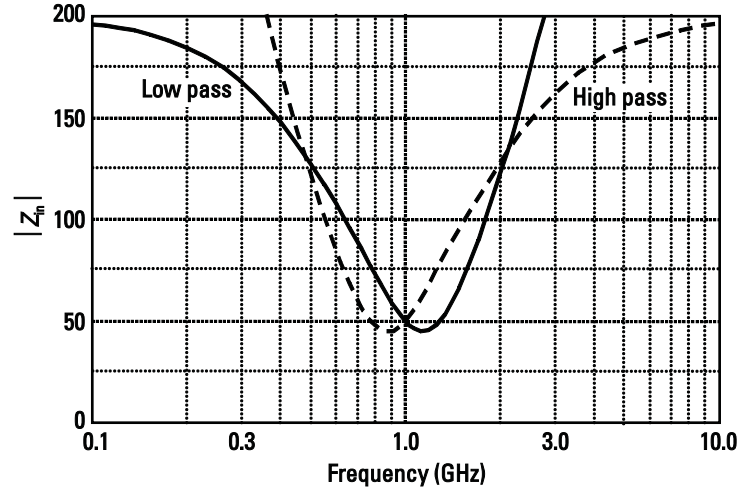

Figure 5.13 Input impedance of lowpass and highpass matching networks.

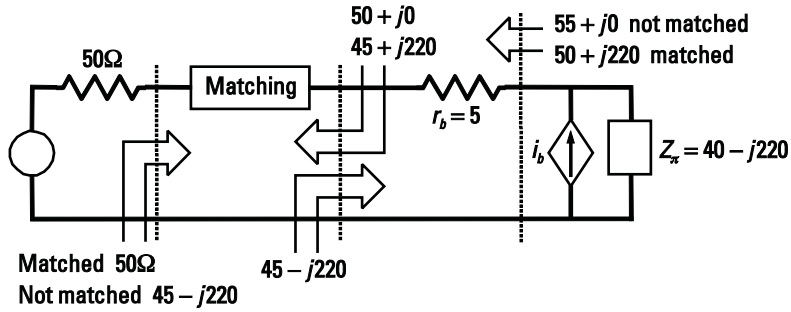

Figure 5.14 Calculation of impact of impedance matching on base shot noise.

5.4 Conversions Between Series and Parallel Resistor-Inductor and Resistor-Capacitor Circuits

Series and parallel resistor-capacitor (RC) and resistor-inductor (RL) networks are widely used basic building blocks of matching networks [1, 2]. In this section, conversions between series and parallel forms of these networks will be discussed. All real inductors and capacitors have resistors (either parasitic or intentional) in parallel or series with them. For the purposes of analysis, it is often desirable to replace these elements with equivalent parallel or series resistors as shown in Figure 5.15.

To convert between series and parallel RC circuits, we first note that the impedance is

$$Z = R_s + \frac{1}{j\omega C_s} \tag{5.7}$$

Converting to an admittance

$$Y = \frac{j\omega C_s + \omega^2 C_s^2 R_s}{1 + \omega^2 C_s^2 R_s^2} \tag{5.8}$$

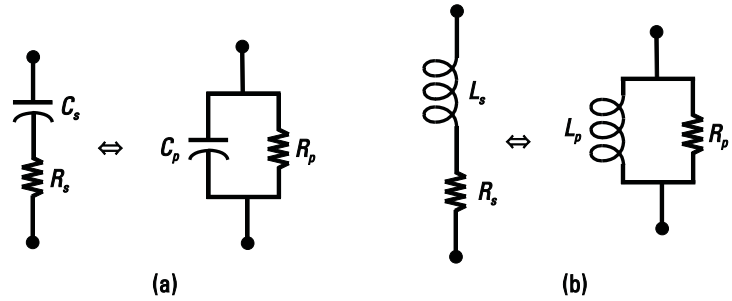

Figure 5.15 Narrowband equivalent models for: (a) a capacitor and series resistor and (b) an inductor and series resistor.

Thus, the inverse of the real part of this equation gives R_p:

$$R_p = \frac{1+\omega^2 C_s^2 R_s^2}{\omega^2 C_s^2 R_s} = R_s(1+Q^2) \tag{5.9}$$

where Q known as the quality factor is defined as before as $|Z_{Im}|/|Z_{Re}|$ where Z_{im} is the imaginary part of Z, and Z_{Re} is the real part of Z. This definition of Q is convenient for the series network, while the equivalent definition of Q as $|Y_{Im}|/|Y_{Re}|$ is more convenient for a parallel network.

The parallel capacitance is thus:

$$C_p = \frac{C_s}{1+\omega^2 C_s^2 R_s^2} = C_s \left(\frac{Q^2}{1+Q^2} \right) \tag{5.10}$$

Similarly for the case of the inductor:

$$R_p = R_s(1+Q^2) \tag{5.11}$$

$$L_p = L_s \left(\frac{1+Q^2}{Q^2} \right) = L_s \left(1+\frac{1}{Q^2} \right) \tag{5.12}$$

For large Q, parallel and series L or C are about the same. Also parallel R is large, while series R is small.

5.5 Tapped Capacitors and Inductors

Another two common basic circuits are shown in Figure 5.16. The figure shows two reactive elements with a resistance in parallel with one of the reactive elements. In this case, the two inductors or two capacitors act to transform the resistance into a higher equivalent value in parallel with the equivalent series combination of the two reactances.

Much as in the previous section the analysis of either Figure 4.16(a) or Figure 4.16(b) begins by finding the equivalent impedance of the network. In the case of Figure 4.16(b) the impedance is given by:

$$Z_{\text{in}} = \frac{j\omega L_1 R + j\omega L_2 R \quad \omega^2 L_1 L_2}{R + j\omega L_2} \tag{5.13}$$

Equivalently the admittance can be found:

$$Y_{\text{in}} = \frac{j\omega R^2 (L_1 + L_2) \quad \omega^2 L_2^2 R + j\omega^3 L_1 {L_2}^2}{\omega^2 R^2 (L_1 + L_2)^2 \quad \omega^4 L_1^2 {L_2}^2} \tag{5.14}$$

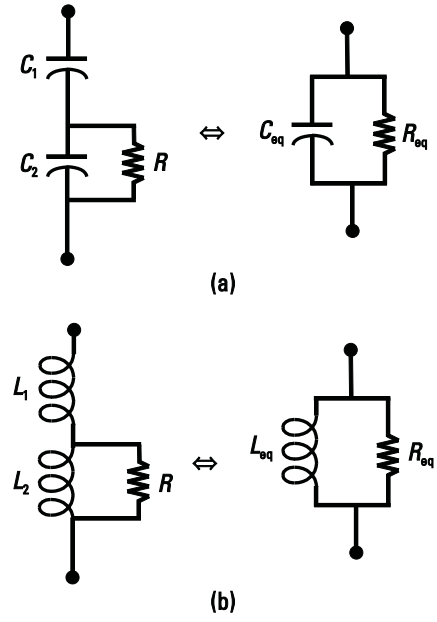

Figure 5.16 Narrowband equivalent models for: (a) a tapped capacitor and resistor and (b) a tapped inductor and resistor.

Thus, the inverse of the real part of this equation gives R_{eq}:

$$R_{eq} = \frac{R^2(L_1 + L_2)^2 \quad \omega^2 L_1^2 L_2^2}{R{L_2}^2} = R\frac{(L_1 + L_2)^2 + \dfrac{L_1^2}{Q_2^2}}{L_2^2} \tag{5.15}$$

where Q_2 is the quality factor of L_2 and R in parallel. As long as Q_2 is large, then a simplification is possible. This is equivalent to stating that the resistance of R is large compared to the impedance of L_2, and the two inductors form a voltage divider.

$$R_{eq} \quad R\left(\frac{L_1 + L_2}{L_2} \div\right)^2 \tag{5.16}$$

The equivalent inductance of the network can be found as well. Again, the inverse of the imaginary part divided by $j\omega$ is equal to the equivalent inductance:

$$L_{eq} = \frac{[R^2(L_1 + L_2)^2 + \omega^2 L_1^2 L_2^2]}{R^2(L_1 + L_2) + \omega^2 L_1 L_2^2} = \frac{(L_1 + L_2)^2 + \dfrac{L_1^2}{Q_2^2}}{L_1 + L_2 + \dfrac{L_1}{Q_2^2}} \tag{5.17}$$

Making the same approximation as before, this simplifies to:

$$L_{eq} \quad L_1 + L_2 \tag{5.18}$$

which is just the series combination of the two inductors if the resistor is absent.

The same type of analysis can be performed on network Figure 4.16(a). In this case

$$R_{\text{eq}} \quad R \quad \frac{C_1 + C_2}{C_1} \div \quad ^2 \tag{5.19}$$

$$C_{\text{eq}} \quad \frac{1}{C_1} + \frac{1}{C_2} \div \quad ^1 \tag{5.20}$$

5.6 The Concept of Mutual Inductance

Any two coupled inductors that affect each other's magnetic fields and transfer energy back and forth form a transformer. How tightly they are coupled together affects how efficiently they transfer energy back and forth. The amount of coupling between two inductors can be quantified by defining a coupling factor k, which can take on any value between one and zero. Another way to describe the coupling between two inductors is with mutual inductance. For two coupled inductors of value L_p and L_s, coupling factor k, and the mutual inductance M as shown in Figure 5.17 are related by

$$k = \frac{M}{\sqrt{L_p L_s}} \tag{5.21}$$

The relationship between voltage and current for two coupled inductors can be written out as follows [3]:

$$\begin{aligned} V_p &= j\omega L_p I_p + j\omega M I_s \\ V_s &= j\omega L_s I_s + j\omega M I_p \end{aligned} \tag{5.22}$$

Note that dots in Figure 5.17 are placed such that if current flows in the indicated direction, then fluxes will be added [4]. Equivalently, if I_p is applied and V_s is 0V, current will be induced opposite to I_s, to minimize the flux.

For transformers, it is necessary to determine where to place the dots. We illustrate this point with Figure 5.18 where voltages V_1, V_2, and V_3 generate flux through a transformer core. The currents are drawn so that the flux is reinforced. The dots are placed appropriately to agree with Figure 5.17.

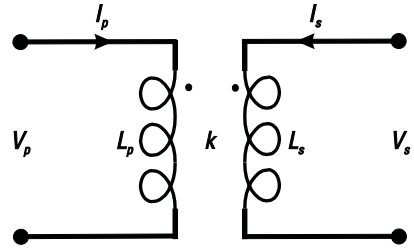

Figure 5.17 A basic transformer structure.

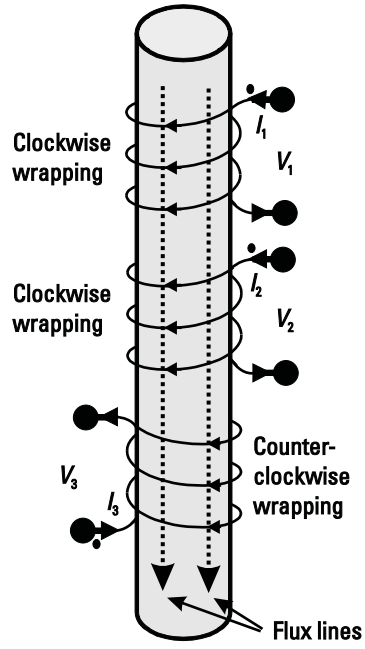

Figure 5.18 Illustration of flux lines and determining correct dot placement.

An equivalent model for the transformer that uses mutual inductance is shown in Figure 5.19. This model can be shown to be valid if two of the ports are connected together as shown in the figure by writing the equations in terms of I_p and I_s, and using the mutual inductance M.

Example 5.5: Equivalent Impedance of Transformer Networks

Referring to the diagram of Figure 5.20, find the equivalent impedance of each structure, noting the placement of the dots.

Solution:
For each structure we apply a test voltage and see what current flows. In the first case, the current flows into the side with the dot of each inductor. In this case, the flux from each structure is added. If we apply a voltage V to circuit on the left in Figure 5.20, then $V/2$ appears across each inductor. Therefore, for each inductor,

$$\frac{V}{2} = j\omega LI + j\omega MI$$

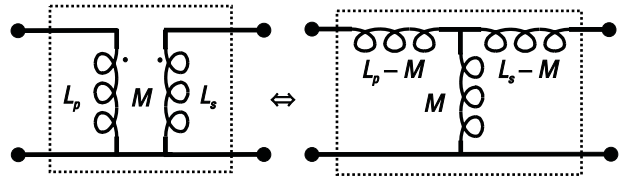

Figure 5.19 Two equivalent models for a transformer.

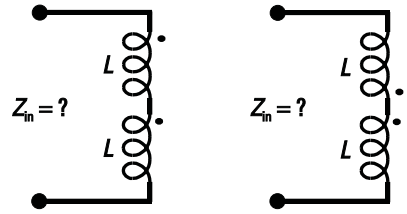

Figure 5.20 Circuits to find the equivalent impedance.

We can solve for the impedance by

$$Z = \frac{V}{I} = 2j\omega(L + M)$$

Thus since $Z = j\omega L_{eq}$, we can solve for L_{eq}:

$$L_{eq} = 2L + 2M$$

In the second case for the circuit on the right in Figure 5.20, the dots are placed in such a way that the flux is reduced. We repeat the analysis:

$$\frac{V}{2} = j\omega LI \quad j\omega MI$$

$$Z = \frac{V}{I} = 2j\omega(L \quad M)$$

$$L_{eq} = 2L \quad 2M$$

Thus, in the first case the inductance reinforces itself, but in the second case it is decreased.

5.7 Matching Using Transformers

Transformers as shown in Figure 5.17 can transform one resistance into another resistance depending on the ratio of the inductance of the primary and the secondary. Assuming that the transformer is ideal (that is, the coupling coefficient k is equal to 1, which means that the coupling of magnetic energy is perfect) and lossless and

$$L_p = NL_s \tag{5.23}$$

then it can be shown from elementary physics that

$$\frac{V_p}{V_s} = \frac{I_s}{I_p} = \sqrt{N} \tag{5.24}$$

Note that here we have defined N as the inductance ratio, but traditionally it is defined as a turns ratio. Since in an integrated circuit turns and inductance are not so easily related, this alternative definition is used.

Now if the secondary is loaded with impedance R_s, then the impedance seen in parallel on the primary side R_p will be:

$$R_p = \frac{V_p}{I_p} = \frac{V_s\sqrt{N}}{\frac{I_s}{\sqrt{N}}} = \frac{V_s}{I_s}N = R_sN \tag{5.25}$$

Thus, the impedance on the primary and secondary are related by the inductance ratio. Therefore, placing a transformer in a circuit provides the opportunity to transform one impedance into another. However, the above expressions are only valid for an ideal transformer where $k = 1$. Also, if the resistor is placed in series with the transformer rather than in parallel with it, then the resistor and inductor will form a voltage divider, modifying the impedance transformation. In order to prevent the voltage divider from being a problem, the transformer must be tuned or resonated with a capacitor so that it provides an open circuit at a particular frequency at which the match is being performed. Thus, there is a trade-off in a real transformer between near-ideal behavior and bandwidth. Of course, the losses in the winding and substrate cannot be avoided.

5.8 Tuning a Transformer

Unlike the previous case where the transformer was assumed to be ideal, in a real transformer there are losses. Since there is inductance in the primary and secondary, this must be resonated out if the circuit is to be matched to a real impedance. To do a more accurate analysis, we start with the equivalent model for the transformer loaded on the secondary with resistance R_L, as shown in Figure 5.21.

Next, we find the equivalent admittance looking into the primary. Through circuit analysis, it can be shown that

$$Y_{\text{in}} = \frac{R_L\omega^2(L_sL_p \quad M^2) \quad j\omega^3(L_sL_p \quad M^2) \quad j\omega R_L{}^2L_p + \omega^2L_sL_pR_L}{\omega^4(L_sL_p \quad M^2)^2 + \omega^2R_s^2L_p{}^2} \tag{5.26}$$

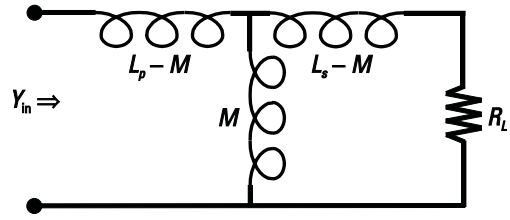

Figure 5.21 Real transformer used to transform one resistance into another.

Taking the imaginary part of this expression, the inductance seen looking into the primary L_{eff} can be found, making use of (5.21) to express the results in terms of the coupling coefficient k:

$$L_{\text{eff}\ \ p} = \frac{\omega^2 L_s^2 L_p (1 \quad k^2)^2 + R_L^2 L_p}{\omega^2 L_s^2 (1 \quad k^2) + R_L^2} \tag{5.27}$$

When $k = 1$, or when $k = 0$, then the inductance is simply L_p. When k has a value between these two limits, then the inductance will be reduced slightly from this value, depending on circuit values. Thus, a transformer can be made to resonate and have a zero reactive component at a particular frequency using a capacitor on either the primary C_p or secondary C_s:

$$\omega_o = \frac{1}{\sqrt{L_{\text{eff}\ \ p} C_p}} = \frac{1}{\sqrt{L_{\text{eff}\ \ s} C_s}} \tag{5.28}$$

where Leff-s is the inductance seen looking into the secondary.

The exact resistance transformation can also be extracted and is given by

$$R_{\text{eff}} = \frac{R_L^2 L_p \quad \omega^2 L_s{}^2 L_p (1 \quad k^2)^2}{R_L L_s k^2} \tag{5.29}$$

Note again that if $k = 1$, then $R_{\text{eff}} = R_L N$ and goes to infinity as k goes to zero.

5.9 The Bandwidth of an Impedance Transformation Network

Using the theory already developed, it is possible to make most matching networks into equivalent parallel or series inductance, resistance, and capacitance (LRC) circuits, such as the one shown in Figure 5.22. The transfer function for this circuit is determined by its impedance, which is given by

$$\frac{V_{\text{out}}(s)}{I_{\text{in}}(s)} = \frac{1}{C} \quad \frac{s}{s^2 + \frac{s}{RC} + \frac{1}{LC} \div} \div \tag{5.30}$$

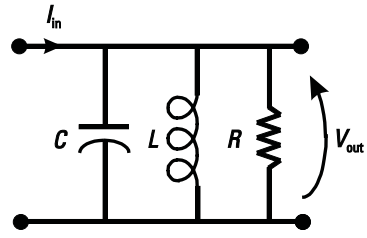

Figure 5.22 An LC resonator with resistive loss.

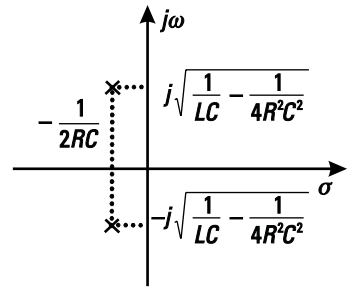

Figure 5.23 Pole plot of an undamped LC resonator.

In general, this second-order transfer function has the form:

$$A(s) = \frac{A_o s}{s^2 + sBW + \omega_o{}^2} \tag{5.31}$$

where

$$BW = \frac{1}{RC} \tag{5.32}$$

and

$$\omega_o = \sqrt{\frac{1}{LC}} \tag{5.33}$$

This is an example of a damped second-order system with poles in the left-hand half plane as shown in Figure 5.23 (provided that R is positive and finite). This system will have a frequency response that is centered on a given resonance frequency ω_o and will fall off on either side of this frequency, as shown in Figure 5.24. The distance on either side of the resonance frequency where the transfer function falls in amplitude by 3 dB is usually defined as the circuit bandwidth. This is the

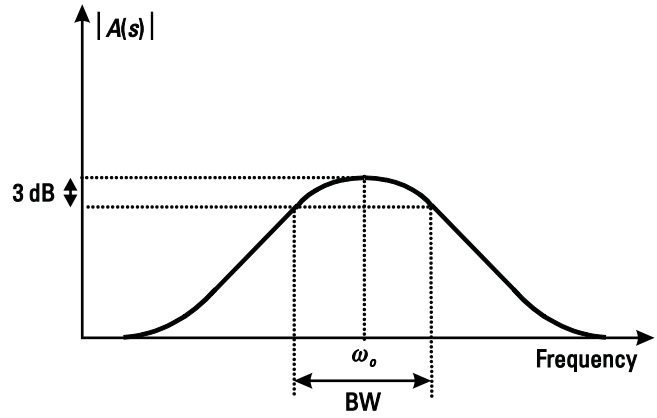

Figure 5.24 Plot of a general second-order bandpass transfer function.

frequency at which the gain of the transfer function is down by 3 dB relative to the gain at the center frequency.

5.10 Quality Factor of an LC Resonator

The quality factor, Q, of an LC resonator is another figure of merit used. It is defined as

$$Q = 2\pi \frac{E_{\text{Stored/Cycle}}}{E_{\text{Lost/Cycle}}} \tag{5.34}$$

This can be used as a starting point to define Q in terms of circuit parameters.

We first note that all the loss must occur in the resistor, because it is the only element present capable of dissipating any energy and the energy dissipated per cycle is:

$$E_{\text{Lost/Cycle}} = \int_0^T \frac{{V_{\text{osc}}}^2 \sin^2(\omega_{\text{osc}} t)}{R} dt = \frac{1}{2} {V_{\text{osc}}}^2 \frac{T}{R} \tag{5.35}$$

Energy is also stored each cycle in the capacitor and the Q is therefore given by:

$$E_{\text{Stored/Cycle}} = \frac{1}{2} C {V_{\text{osc}}}^2 \qquad Q = 2\pi \frac{CR}{T} = CR\omega_{\text{osc}} = R\sqrt{\frac{C}{L}} \tag{5.36}$$

Another definition of Q that is particularly useful is [5]:

$$Q = \frac{\omega_o}{2} \left| \frac{d\phi}{d\omega} \right| \tag{5.37}$$

where ϕ is the phase of the resonator and $d\phi/d\omega$ is the rate of change of the phase transfer function with respect to frequency. This can be shown to give the same value in terms of circuit parameters as (5.34).

The Q of a resonator can also be related to its center frequency and bandwidth, noting that:

$$Q = R\sqrt{\frac{C}{L}} = \frac{RC}{\sqrt{LC}} = \frac{\omega_o}{\text{BW}} \tag{5.38}$$

Example 5.6: Matching a Transistor Input with a Transformer

A circuit has an input that is made up of a 1-pF capacitor in parallel with a 200 resistor. Use a transformer with a coupling factor of 0.8 to match it to a source resistance of 50 . The matching circuit must have a bandwidth of 200 MHz and the circuit is to operate at 2 GHz.

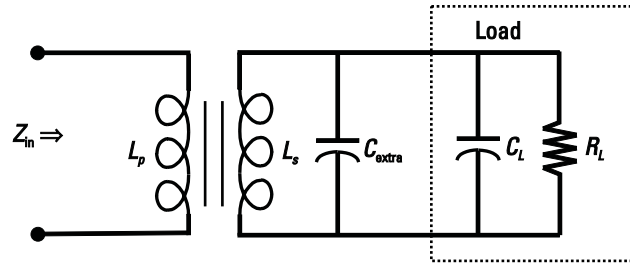

Figure 5.25 Transformer matching network used to match the input of a transistor.

Solution:
The matching circuit will look much like that shown in Figure 5.25. We will use the secondary of the transformer as a resonant circuit so that there will be no reactance at 2 GHz. We first add capacitance in parallel with the input capacitance so that the circuit will have the correct bandwidth. Using (5.32),

$$C_{\text{total}} = \frac{1}{RBW} = \frac{1}{(100\quad)(2\pi\times 200\text{ MHz})} = 7.96\text{ pF}$$

Note that the secondary "sees" 100 total as there is 200 from the load and 200 from the source resistance. This means that C_{extra} must be 6.96 pF. Now to resonate at 2 GHz, this means that the secondary of the transformer must have an inductance of:

$$L_s = \frac{1}{\omega_o{}^2 C_s} = \frac{1}{(2\pi\times 2\text{ GHz})^2(7.96\text{ pF})} = 0.8\text{ nH}$$

Now we must set the inductance ratio to turn 200 into 50 :

$$L_p = \frac{R_{\text{eff}} R_L L_s k^2}{R_L{}^2\quad \omega_o{}^2 L_s{}^2(1\quad k^2)^2} = \frac{50\quad\times 200\quad\times 0.8\text{ nH}\times(0.8)^2}{(200\quad)^2\quad(2\pi\times 2\text{ GHz})^2(0.8\text{ nH})^2(1\quad 0.8^2)^2} = 0.13\text{ nH}$$

Example 5.7: Matching Using a Two Stage "Ell" Network

Match 200 to 50 using an "ell" matching network. Do it first in one step, then do it in two steps matching it first to 100 . Compare the bandwidth of the two matching networks.

Solution:
Figure 5.26 illustrates matching done in one step (with movement from a to b to c) versus matching done in two steps (with movement from a to d to c). One-step matching was previously shown in Example 5.4 and Figure 5.11. Two-step matching calculations are also straightforward with an ell network converting from 200 to 100 , then another ell network converting from 100 to 50 . The resulting network is shown in Figure 5.27.

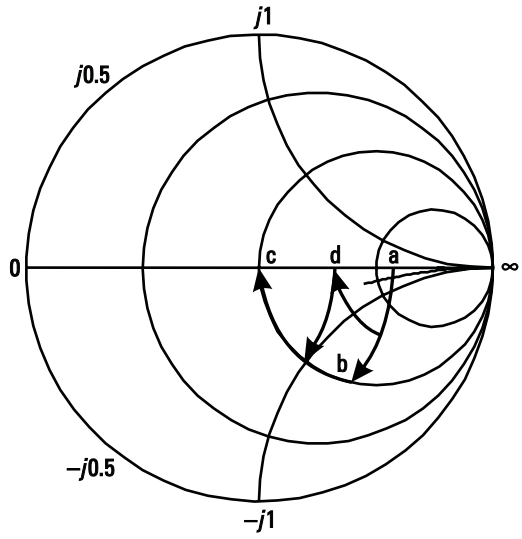

Figure 5.26 Smith chart illustration of one-step versus two-step matching.

A comparison of frequency response (Figure 5.28) clearly shows the bandwidth broadening effect of matching in two steps. To quantify the effect, the magnitude of the input impedance is shown in Figure 5.29.

5.11 Broadband Impedance Matching

In many cases it may be necessary to match a circuit over a band of frequencies, which cannot be considered narrow with respect to the center frequency. In this case, more than two matching elements are usually required. One technique for matching makes use of a bandpass ladder filter [6, 7] in which a matching network is constructed as shown in Figure 5.30. Using this technique, the device that is being matched is designed to have an input impedance where the real part is designed to be equal to the source impedance. While this may seem very restrictive, a technique for making devices have a desired resistive component will be studied in Chapter 7. In general, the device will still have a reactive component that will make its impedance frequency dependent. A ladder of series and parallel LC networks is added to the input port of the device so that

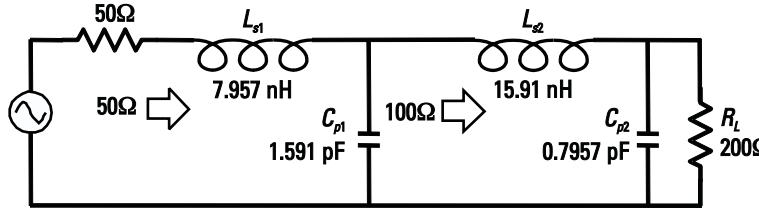

Figure 5.27 Circuit for two-step matching.

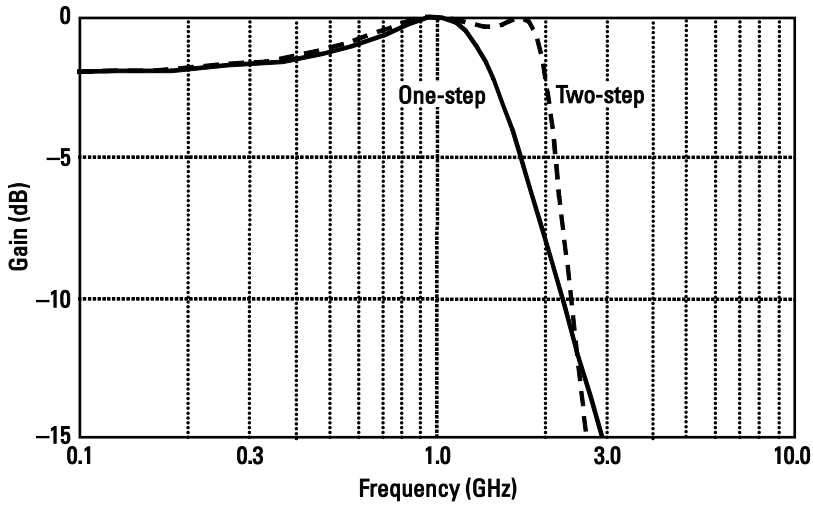

Figure 5.28 Frequency response for one-step and two-step matching.

$$
\begin{aligned}
L_{1,3} &= \frac{R_s}{\omega_H} \\
C_2 &= \frac{1}{\omega_H R_s} \\
L_2 &= \frac{R_s}{\omega_L} \\
C_3 &= \frac{1}{\omega_L R_s}
\end{aligned}
\tag{5.39}
$$

where ω_H and ω_L are the upper and lower frequencies over which the circuit is to be matched. Note that the circuit in Figure 5.30 is shown with three stages; how-

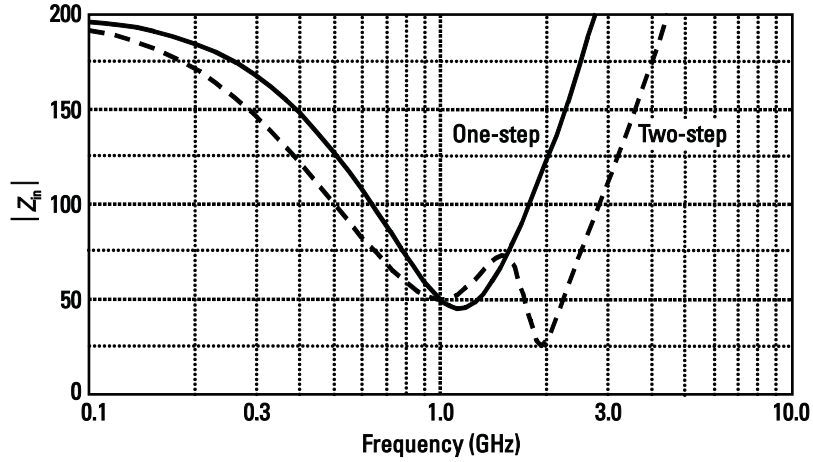

Figure 5.29 Input Impedance for one-step and two-step matching.

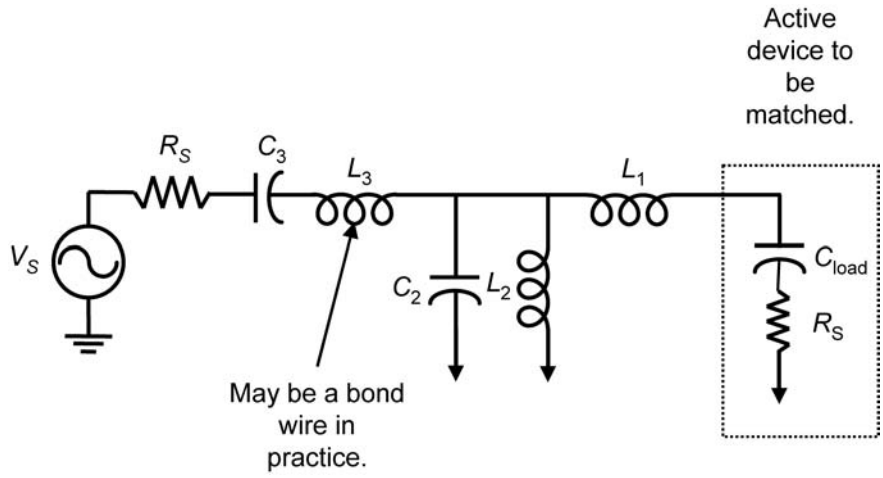

Figure 5.30 A broadband input impedance matching network.

ever, an arbitrary number of ladder stages may be used in theory. More stages can cause additional loss if the components are less than ideal, but will have the added advantage that any process tolerance causing a change in any one value will have a lesser impact on the overall circuit matching.

Example 5.8: A UWB Matching Circuit

Design a matching network to match an active device from 3–5 GHz. The device is sized so that it has an input impedance of 800 fF in series with 50 .

Solution:
The values for the ladder can be chosen with the help of (5.39) such that:

$$L_{1,3} = \frac{50}{2\pi \times 5\ \text{GHz}} = 1.59\ \text{nH}$$

$$C_2 = \frac{1}{2\pi \times 5\ \text{GHz} \times 50} = 636.6\ \text{fF}$$

$$L_2 = \frac{50}{2\pi \times 3\ \text{GHz}} = 2.65\ \text{nH}$$

$$C_3 = \frac{1}{2\pi \times 3\ \text{GHz} \times 50} = 1.06\ \text{pF}$$

Note that ideally C_{Load} should also be 1.06 pF; a capacitor could be added to fix this. However, in an active circuit this might have other undesirable consequences and a slight mismatch will likely only marginally impact the design. Constructing the rest of the ladder, the reflection coefficient is shown in Figure 5.31, which shows that the circuit is matched to better than 10 dB over the required bandwidth. A reflection coefficient of 10 dB represents only a small loss of power transfer and in practice is considered a good match.

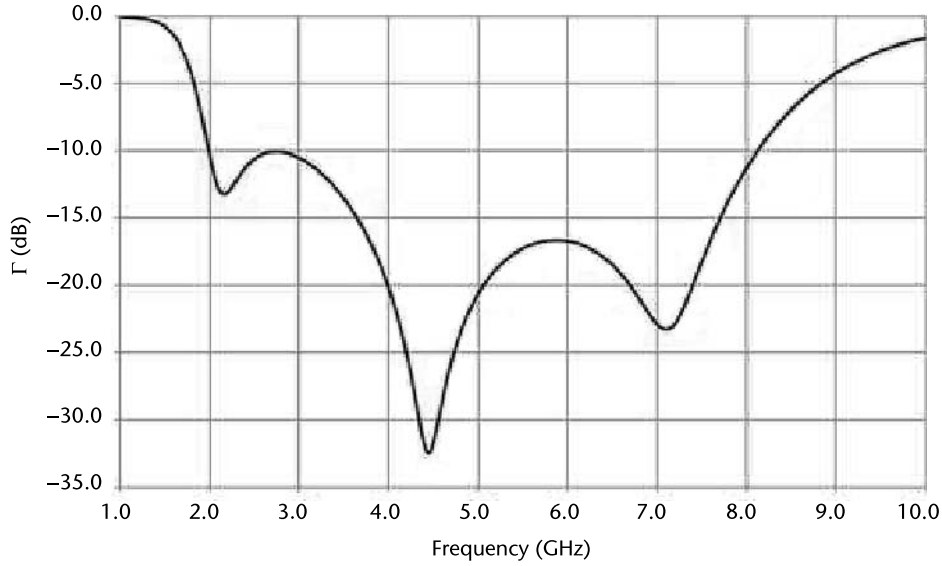

Figure 5.31 The reflection coefficient of a device matched from 3–5 GHz.

5.12 Transmission Lines

When designing circuits on-chip, often transmission line effects can be ignored, but at chip boundaries they are very important. Transmission lines have effects that must be considered at these interfaces in order to match the input or output of an RFIC. As already discussed, transmission lines have a characteristic impedance, and when they are loaded with an impedance different from this characteristic impedance, they cause the impedance looking into the transmission line to change with distance. If the transmission line, such as shown in Figure 5.32, is considered lossless, then the input impedance at any distance d from the load is given by [8]

$$Z_{\text{in}}(d) = Z_o \frac{Z_L + jZ_o \tan \frac{2\pi}{\lambda} d\div}{Z_o + jZ_L \tan \frac{2\pi}{\lambda} d\div} \tag{5.40}$$

where λ is one wavelength of an EM wave at the frequency of interest in the transmission line. A brief review of how to calculate λ will be given in Chapter 6. Thus,

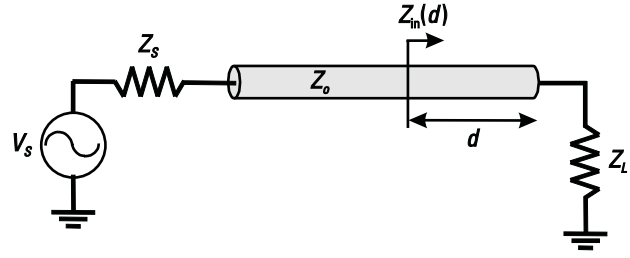

Figure 5.32 Impedance seen moving down a transmission line.

the impedance looking into the transmission line is periodic with distance. It can be shown from (5.40) that for each distance λ traveled down the transmission line, the impedance makes two clockwise rotations about the center of the Smith chart.

Transmission lines can also be used to synthesize reactive impedances. Note that if Z_L is either an open or short circuit, then by making the transmission line an appropriate length, any purely reactive impedance can be realized. These types of transmission lines are usually referred to as *open-* and *short-circuit stubs*. This topic will not be addressed further because the size of transmission line matching elements over much of the frequency range of RFICs is very large compared with the relatively small size of lumped inductors, capacitors, and transformers.

5.13 *S, Y,* and *Z* Parameters

S, *Y*, and *Z* parameters (scattering, admittance, and impedance parameters, respectively) are widely used in the analysis of RF circuits. For RF measurements (for example, with a network analyzer), typically S-parameters are used. These may be later converted to *Y* or *Z* parameters in order to perform certain analyses. In this section, *S*-parameters and conversions to other parameters will be described.

S-parameters are a way of calculating a two-port network in terms of incident and reflected (or scattered) power. Referring to Figure 5.33, assume that port 1 is the input, a_1 is the input wave, b_1 is the reflected wave, and b_2 is the transmitted wave. We note that if a transmission line is terminated in its characteristic impedance, then the load absorbs all incident power traveling along the transmission line, and there is no reflection.

The *S*-parameters can be used to describe the relationship between these waves as follows:

$$b_1 = S_{11}a_1 + S_{12}a_2 \tag{5.41}$$

$$b_2 = S_{22}a_2 + S_{21}a_1 \tag{5.42}$$

This can also be written in a matrix as:

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad [b] = [S][a] \tag{5.43}$$

Thus, *S*-parameters are reflection or transmission coefficients and are usually normalized to a particular impedance. S_{11}, S_{22}, S_{21}, and S_{12} will now each be defined.

$$S_{11} = \left.\frac{b_1}{a_1}\right|_{a_2=0} \tag{5.44}$$

Figure 5.33 General two-port system with incident and reflected waves.

S_{11} is the input reflection coefficient, measured with the output terminated with Z_o. This means the output is matched and all power is transmitted into the load; thus a_2 is zero.

$$S_{21} = \left.\frac{b_2}{a_1}\right|_{a_2=0} \tag{5.45}$$

S_{21}, the forward transmission coefficient, is also measured with the output terminated with Z_o. S_{21} is equivalent to gain.

$$S_{22} = \left.\frac{b_2}{a_2}\right|_{a_1=0} \tag{5.46}$$

S_{22} is the output reflection coefficient, measured by applying a source at the output and with the input terminated with Z_o.

$$S_{12} = \left.\frac{b_1}{a_2}\right|_{a_1=0} \tag{5.47}$$

S_{12} is the reverse transmission coefficient, measured with the input terminated with Z_o.

In addition to *S*-parameters, there are many other parameter sets that can be used to characterize a two-port network. Since engineers are used to thinking in terms of voltages and currents, another popular set of parameters is *Z*- and *Y*-parameters as shown in Figure 5.34. *Y*- and *Z*-parameters can be used to describe the relationship between voltages and currents as follows:

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} \tag{5.48}$$

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{5.49}$$

It is also useful to be able to translate from one set of these parameters to the other. These relationships are well known, and are summarized in Table 5.3.

Microwave transistors or amplifiers are often completely (and exclusively) characterized with *S*-parameters. For radio frequency integrated circuits, detailed transistor models are typically used, which allow the designer (with the help of the simulator) to design circuits. The models and simulators can be used to find *S*-

Figure 5.34 General two-port system with input and output currents and voltages.

Table 5.3 Relationships Between Different Parameter Sets

	S	Z	Y
S_{11}	S_{11}	$\frac{(Z_{11} - Z_o)(Z_{22} + Z_o) - Z_{12}Z_{21}}{(Z_{11} + Z_o)(Z_{22} + Z_o) - Z_{12}Z_{21}}$	$\frac{Y_o - Y_{11}(Y_{22} + Y_o) + Y_{12}Y_{21}}{(Y_{11} + Y_o)(Y_{22} + Y_o) - Y_{12}Y_{21}}$
S_{12}	S_{12}	$\frac{2Z_{12}Z_o}{(Z_{11} + Z_o)(Z_{22} + Z_o) - Z_{12}Z_{21}}$	$\frac{2Y_{12}Y_o}{(Y_{11} + Y_o)(Y_{22} + Y_o) - Y_{12}Y_{21}}$
S_{21}	S_{21}	$\frac{2Z_{21}Z_o}{(Z_{11} + Z_o)(Z_{22} + Z_o) - Z_{12}Z_{21}}$	$\frac{2Y_{21}Y_o}{(Y_{11} + Y_o)(Y_{22} + Y_o) - Y_{12}Y_{21}}$
S_{22}	S_{22}	$\frac{(Z_{11} + Z_o)(Z_{22} - Z_o) - Z_{12}Z_{21}}{(Z_{11} + Z_o)(Z_{22} + Z_o) - Z_{12}Z_{21}}$	$\frac{(Y_o + Y_{11})(Y_0 - Y_{22}) + Y_{12}Y_{21}}{(Y_{11} + Y_o)(Y_{22} + Y_o) - Y_{12}Y_{21}}$
Z_{11}	$Z_o\frac{(1 + S_{11})(1 - S_{22}) + S_{12}S_{21}}{(1 - S_{11})(1 - S_{22}) - S_{12}S_{21}}$	Z_{11}	$\frac{Y_{22}}{Y_{11}Y_{22} - Y_{12}Y_{21}}$
Z_{12}	$Z_o\frac{2S_{12}}{(1 - S_{11})(1 - S_{22}) - S_{12}S_{21}}$	Z_{12}	$\frac{Y_{12}}{Y_{11}Y_{22} - Y_{12}Y_{21}}$
Z_{21}	$Z_o\frac{2S_{21}}{(1 - S_{11})(1 - S_{22}) - S_{12}S_{21}}$	Z_{21}	$\frac{Y_{21}}{Y_{11}Y_{22} - Y_{12}Y_{21}}$
Z_{22}	$Z_o\frac{(1 + S_{22})(1 - S_{11}) + S_{12}S_{21}}{(1 - S_{11})(1 - S_{22}) - S_{12}S_{21}}$	Z_{22}	$\frac{Y_{11}}{Y_{11}Y_{22} - Y_{12}Y_{21}}$
Y_{11}	$Y_o\frac{(1 + S_{22})(1 - S_{11}) + S_{12}S_{21}}{(1 + S_{11})(1 + S_{22}) - S_{12}S_{21}}$	$\frac{Z_{22}}{Z_{11}Z_{22} - Z_{12}Z_{21}}$	Y_{11}
Y_{12}	$Y_o\frac{2S_{12}}{(1 + S_{11})(1 + S_{22}) - S_{12}S_{21}}$	$\frac{Z_{12}}{Z_{11}Z_{22} - Z_{12}Z_{21}}$	Y_{12}
Y_{21}	$Y_o\frac{2S_{21}}{(1 + S_{11})(1 + S_{22}) - S_{12}S_{21}}$	$\frac{Z_{21}}{Z_{11}Z_{22} - Z_{12}Z_{21}}$	Y_{21}
Y_{22}	$Y_o\frac{(1 - S_{22})(1 + S_{11}) + S_{12}S_{21}}{(1 + S_{11})(1 + S_{22}) - S_{12}S_{21}}$	$\frac{Z_{11}}{Z_{11}Z_{22} - Z_{12}Z_{21}}$	Y_{22}

parameters, which can be used with the well-known microwave techniques to find maximum gain, optimal noise figure, stability, and so forth. However, the simulators, which use the models to generate the *S*-parameters, can be used directly to find maximum gain, optimal noise figure, and stability, without the need to generate a list of *S*-parameters. However, it is worthwhile to be familiar with these design techniques since they can give insight into circuit design, which can be of much more value than simply knowing the location of the "simulate" button.

References

[1] Krauss, H. L., C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, New York: John Wiley & Sons, 1980.

[2] Smith, J. R., *Modern Communication Circuits*, 2nd ed., New York: McGraw-Hill, 1998.

[3] Irwin, J. D., *Basic Engineering Circuit Analysis*, New York: Macmillan, 1993.

[4] Sadiku, M. N. O., *Elements of Electromagnetics*, 2nd ed., Fort Worth, TX: Sanders College Publishing, 1994.

[5] Razavi, B., "A Study of Phase Noise in CMOS Oscillators," *IEEE J. Solid-State Circuits*, Vol. 31, March 1996, pp. 331–343.

[6] Ismail, A., and A. Abidi, "A 3–10-GHz Low-Noise Amplifier with Wideband LC-Ladder Matching Network," *IEEE J. Solid-State Circuits*, Vol. 39, December 2004, pp. 2269–2277.

[7] Bevilacqua, A., and A. Niknejad, "An Ultrawideband CMOS Low-Noise Amplifier for 3.1–10.6-GHz Wireless Receivers," *IEEE J. Solid-State Circuits*, Vol. 39, December 2004, pp. 2259–2268.

[8] Pozar, D. M., *Microwave Engineering*, 2nd ed., New York: John Wiley & Sons, 1998.

CHAPTER 6

The Use and Design of Passive Circuit Elements in IC Technologies

6.1 Introduction

In this chapter, passive circuit elements will be discussed. First, metalization and back-end processing (away from the silicon) in integrated circuits will be described. This is the starting point for many of the passive components. Then, design, modeling, and use of passive components will be discussed. These components are interconnect lines, inductors, capacitors, transmission lines, and transformers. Finally, there will be a discussion of the impact of packaging.

Passive circuit elements such as inductors and capacitors are necessary components in RF circuits, but these components often limit performance, so it is worthwhile to study their design and use. For example, inductors have many applications in RF circuits, as summarized in Table 6.1. An important property of the inductor is that it can simultaneously provide low impedance to dc while providing nite ac impedance. In matching circuits or tuned loads, this allows active circuits to be biased at the supply voltage for maximum linearity. However, inductors are lossy, resulting in increased noise when used in an LNA or oscillator. When used in a power ampli er, losses in inductors can result in decreased ef ciency. Also, substrate coupling is a serious concern because of the typically large physical dimensions of the inductor.

6.2 The Technology Back End and Metalization in IC Technologies

After all the front-end processing is complete, the active devices are connected using metal (the back end), which is deposited above the transistors as shown in Figure 6.1. The metals must be placed in an insulating layer of silicon dioxide (SiO_2), to prevent different layers of metal from shorting with each other. Most processes have several layers of metal in their back end. These metal layers can also be used to build capacitors, inductors, and even resistors.

The bottom metal is typically connected to the front end with tungsten plugs, which are highly resistive. However, unlike aluminum, gold, or copper, this metal has the property that it will not diffuse into the silicon. When metals such as copper diffuse into silicon, they cause the junction to leak, seriously impairing the performance of transistors. A contact layer is used to connect this tungsten layer to the active circuitry in the silicon. Higher levels of metal can be connected to adjacent layers using conductive plugs that are commonly called vias. Whereas metal can be made in almost any shape desired by the designer, the vias are typically limited to

Table 6.1 Applications and Bene ts of Inductors

Circuit	Application	Bene t
LNA	Input match, degeneration	Simultaneous power and noise matching, improved linearity
	Tuned load	Biasing for best linearity, ltering, less problems with parasitic capacitance now part of resonant circuit
Mixer	Degeneration	Increased linearity, reduced noise
Oscillator	Resonator	Sets oscillating frequency, high Q circuit results in reduced power requirement, lower phase noise
Power ampli er	Matching loads	Maximize voltage swings, higher ef ciency due to swing, (inductor losses reduce the ef ciency)

a standard square size. However, it is possible to use arrays of vias to reduce the resistance.

Higher metal layers are often made out of aluminum or copper, as they are much less resistive than tungsten. The top level of metal will often be made much thicker than the lower levels to provide a low-resistance routing option. However, the lithography for this layer may be much coarser than that of underlying layers. Thus, the top layers can accommodate a lower density of routing lines.

6.3 Sheet Resistance and the Skin Effect

All conductive materials can be characterized by their resistivity or their conductivity . These two quantities are related by

$$\rho = \frac{1}{\sigma} \tag{6.1}$$

Resistivity is expressed as ohm-meters (m). Knowing the geometry of a metal and its resistivity is enough to estimate the resistance between any two points con-

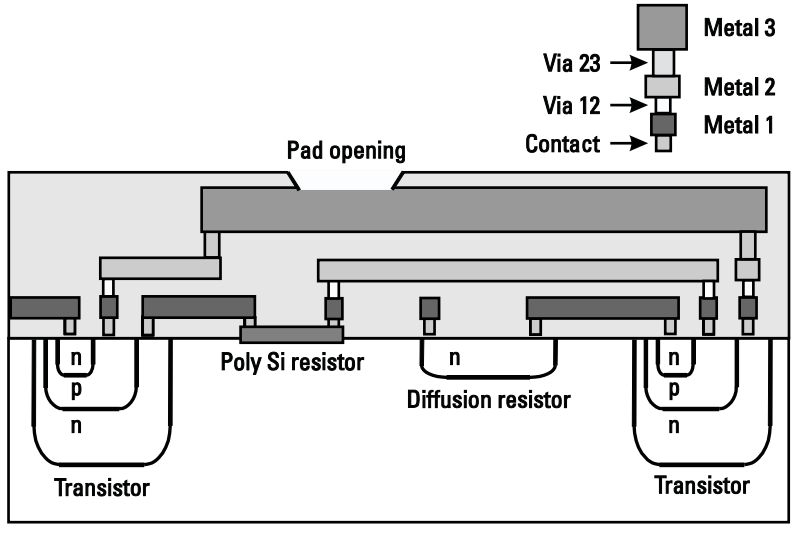

Figure 6.1 Cross section of a typical bipolar back-end process.

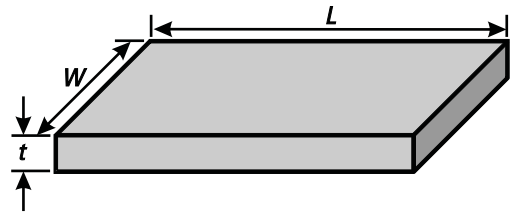

Figure 6.2 Rectangular conductor with current owing in the direction of L.

nected by the metal. As an example, consider the conductor shown in Figure 6.2. To nd the resistance along its length, divide the resistivity of the metal by the cross sectional area and multiply by the length.

$$R = \frac{\rho L}{Wt} \tag{6.2}$$

Often in IC technologies, sheet resistance is used instead of resistivity. Sheet resistance is given by

$$\rho_s = \frac{\rho}{t} = R \left(\frac{W}{L} \right) \tag{6.3}$$

Given the sheet resistance, typically expressed as ohms per square ( ), to nd resistance, multiply by the number of squares between input and output. That is to say, for every distance traveled down the conductor equal to its width , a square has been de ned. If the conductor has a length equal to twice its width, then it is two squares long.

As the metal gets thicker, the resistance of the line decreases. However, the lithography of the process becomes harder to control. Thick metal lines close to one another also suffer from capacitance between the two adjacent sidewalls. At high frequencies, another effect comes into play as well. EM waves suffer attenuation as they enter a conductor, so as the frequency approaches the gigahertz range, the distance that the waves can penetrate becomes comparable to the size of the metal line. The result is that the current becomes concentrated around the outside of the conductor with very little owing in the center. The depth at which the magnitude of the EM wave is decreased to 36.8% ( ) of its intensity at the surface is called the skin depth of the metal. The skin depth is given by

$$\delta = \sqrt{\frac{\rho}{\pi f \mu}} \tag{6.4}$$

where f is the frequency, and is the permeability of the metal. Table 6.2 shows the skin depth of some common metals over the frequency band of interest.

Since most of the applications lie in the 900-MHz to 60-GHz band, it is easy to see that making lines much thicker than about 4µm will lead to diminishing returns. Going any thicker will yield little advantage at the frequencies of interest because the center of the conductor will form a “dead zone” where little current will ow anyway.

Table 6.2 Skin Depth of Various Metals at Various Frequencies

Metal	(μW×cm)	500 MHz	1 GHz	2 GHz	5 GHz	10 GHz
Gold	2.44	3.5 μm	2.5 μm	1.8 μm	1.1 μm	0.79 μm
Tungston	5.49	5.3 μm	3.7 μm	2.6 μm	1.7 μm	1.2 μm
Aluminum	2.62	3.6 μm	2.6 μm	1.8 μm	1.2 μm	0.82 μm
Copper	1.72	3.0 μm	2.1 μm	1.5 μm	0.93 μm	0.66 μm
Silver	1.62	2.9 μm	2.0 μm	1.4 μm	0.91 μm	0.64 μm
Nickel	6.90	5.9 μm	4.2 μm	3.0 μm	1.9 μm	1.3 μm

Example 6.1: Effect of Skin Depth on Resistance

A rectangular aluminum line has a width of 20 μm, a thickness of 3μm, and a length of 100 μm. Compute the resistance of the line at dc and at 5 GHz assuming that all the current flows in an area one skin depth from the surface. Assume that aluminum has a resistivity of 3 μWcm. Note that there are more complex equations, which describe the resistance due to skin effects, especially for circular conductors [1]; however, the simple estimate used here will illustrate the nature of the skin effect.

Solution:
The dc resistance is given by:

$$R = \frac{\rho L}{Wt} = \frac{3\ \mu\ \text{cm} \times 100\ \mu\text{m}}{20\ \mu\text{m} \times 3\ \mu\text{m}} = 50\ \text{m}$$

At 5 GHz, the skin depth of aluminum is:

$$\delta = \sqrt{\frac{\rho}{\pi f \mu}} = \sqrt{\frac{3\ \mu\ \text{cm}}{\pi \times 5\ \text{GHz} \times 4\pi \times 10^{-7}\ \frac{N}{A^2}}} = 1.23\ \mu\text{m}$$

We now need to modify the original calculation and divide by the useful cross-sectional area rather than the actual cross-sectional area.

$$R = \frac{\rho L}{Wt\ \ (W\ \ 2\delta)(t\ \ 2\delta)} = \frac{3\ \mu\ \text{cm} \times 100\ \mu\text{m}}{20\ \mu\text{m} \times 3\ \mu\text{m}\ \ 17.5\ \mu\text{m} \times 0.54\ \mu\text{m}} = 59.3\ \text{m}$$

This is almost a 20% increase. Thus, while we may be able to count on process engineers to give us thicker metal, this may not solve all our problems.

6.4 Parasitic Capacitance

Metal lines, as well as having resistance associated with them, also have capacitance. Since the metal in an IC technology is embedded in an insulator over a silicon substrate with low resistivity, the metal trace and the substrate (acting as a lossy conductor) form a parallel-plate capacitor. The parasitic capacitance of a metal line can be approximated by

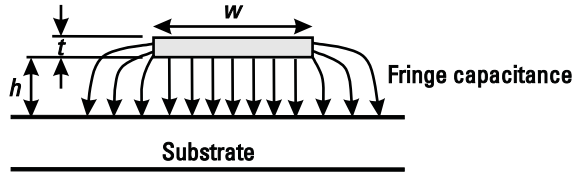

Figure 6.3 Electric eld lines showing the effect of fringing capacitance.

$$C = \frac{\varepsilon_o \varepsilon_r A}{h} \tag{6.5}$$

where A is the area of the trace and h is the distance to the substrate.

Since metal lines in ICs can often be quite narrow, the fringing capacitance can be important, as the electric elds are not con ned to being below the conductor, as shown in Figure 6.3.

For a long line, the capacitance per unit length, taking into account fringing capacitance, can be determined from [2]

$$C = \varepsilon_o \varepsilon_r \left[\frac{W}{h} + 0.77 + 1.06 \left(\frac{W}{h} \div \right)^{\frac{1}{4}} + 1.06 \left(\frac{t}{h} \div \right)^{\frac{1}{2}} \right] \tag{6.6}$$

We note that the terms in the square brackets are unitless, the nal capacitance is given by the units of ε_o (ε_o is 8.85×10^{-12} F/m, ε_r for SiO_2 is 3.9). The rst term accounts for the bottom-plate capacitance while the other three terms account for fringing capacitance. As will be seen in Example 6.2, wider lines will be less affected by fringing capacitance.

Note there is also capacitance between lines vertically and horizontally. A rough estimate of capacitance would be obtained by using the parallel-plate capacitance formula; however, this omits the fringing capacitance, so it would be an underestimate. So, what is the effect of such capacitance? For one, it can lead to crosstalk between parallel lines, or between lines that cross over. For parallel lines, crosstalk can be reduced by further separation, or by placing a ground line between the two signal carrying lines.

Example 6.2: Calculation of Capacitance

Calculate bottom plate capacitance and fringing capacitance for a 1-poly, 4-metal process with distances to substrate and conductor thickness as given in the rst two rows of Table 6.3. Calculate for a metal width of 1 µm and 50 µm.

Solution:

Bottom plate capacitance can be estimated from (6.5), which is equivalent to the rst term in (6.6). Total capacitance can be calculated from (6.6) and the difference attributed to fringing capacitance. Results are shown for the 1µm line and the 50-µm line in Table 6.3. It can be seen that bottom plate capacitance is a very poor estimate of total capacitance for a 1-µm line. When calculated for a 50-µm width, the

Table 6.3 Capacitance for a Line with a Width of 1 µm and 50 µm

	Poly	Metal 1	Metal 2	Metal 3	Metal 4
Height above substrate h (µm)	0.4	1.0	2.5	4.0	5.0
Conductor thickness t (µm)	0.4	0.4	0.5	0.6	0.8
Bottom-plate capacitance (aF/µm^2)	86.3	34.5	13.8	8.6	6.9
Total capacitance (aF/µm^2) (1µ line)	195.5	120.8	85.8	75.2	72.6
Total capacitance (aF/µm^2) (50µ line)	90.0	37.5	16.2	10.8	9.0

bottom plate capacitance and the total capacitance are in much closer agreement. This example clearly shows the inaccuracies inherent in a simple calculation of capacitance. Obviously, it is essential that layout tools have the ability to determine parasitic capacitance accurately.

6.5 Parasitic Inductance

As well as capacitance to the substrate, metal lines in ICs also have inductance. The current owing in the line will generate magnetic eld lines as shown in Figure 6.4. Note that the Xs indicate current ow into the page.

For a at trace of width w, and a distanceh above a ground plane, an estimate for inductance in nH/mm is [3]:

$$L \quad \frac{1.6}{K_f} \times \frac{h}{w} \tag{6.7}$$

Here K_f is the fringe factor, which can be approximated as:

$$K_f \quad 0.72 \times \frac{h}{w} + 1 \tag{6.8}$$

Example 6.3: Calculation of Inductance

Calculate the inductance per unit length for traces with ah/w of 0.5, 1, and 2.

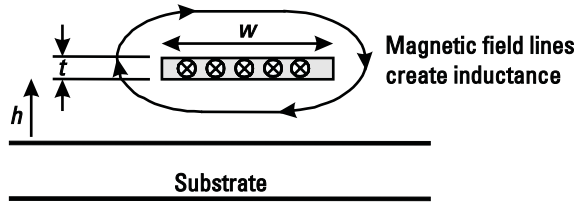

Figure 6.4 Magnetic eld lines around an IC line carrying current.

Solution:
Application of (6.7) and (6.8) show that for h/w is 0.5, 1, 2, the resultant L is 0.59, 0.93, 1.31 nH/mm. A typical rule of thumb is that bond wires have an inductance of 1 nH/mm. This rule of thumb can be seen to also apply approximately to a metal line on-chip.

6.6 Current Handling in Metal Lines

As one can imagine, there is a nite amount of current that can be forced down an IC interconnect before it fails. However, even if the line refrains from exploding this does not necessarily mean that the current is acceptable for long-term reliability. The main mechanism for loss of reliability is metal migration. Metal migration is related to the level of dc current, and this information is used to specify current limits in an IC. To explain metal migration, consider that normally the diffusion process is random, but with dc current, metal atoms are bombarded more from one side than from the other. This causes the movement of metal atoms, which is referred to as metal migration. Suf cient movement in the metal can result in gaps or open circuits appearing in metal and subsequent circuit failure. Any defects or grain boundaries can make the problems worse.

The maximum allowable current in a metal line also depends on the material. For example, aluminum, though lower in resistance, is worse than tungsten for metal migration, due to its much lower melting temperature. Thus, even though there is less energy dissipated per unit length in aluminum, it is less able to handle that energy dissipation.

For 1-µm-thick aluminum, a typical value for maximum current would be 1 mA of dc current for every micrometer of metal width. Similarly, a 2-µm-thick aluminum line would typically be able to carry 2 mA of dc current per micrometer of metal width. The ac component of the current can be larger (a typical factor of 4 is often used). We note that other metals like copper and gold are somewhat lower in resistance than aluminum; however, due to better metal migration properties they can handle more current than aluminum.

Example 6.4: Calculating Maximum Line Current

If a line carries no dc current, but has a peak ac current of 500 mA, a 1µm-thick metal line would need to be about 500 mA/4 mA/µm = 125 µm wide. However, if the dc current is 500 mA, and the peak ac current is also 500 mA (i.e., 500 mA± 500 mA), then the 500-µm-wide line required to pass the dc current is no longer quite wide enough. To cope with the additional ac current, another 125µm is required for a total width of 625 µm.

6.7 Poly Resistors and Diffusion Resistors

Poly resistors are made out of conductive polycrystalline silicon that is directly on top of the silicon front end. Essentially this layer acts like a resistive metal line. Typically these layers have a resistivity in the 10W/F range.

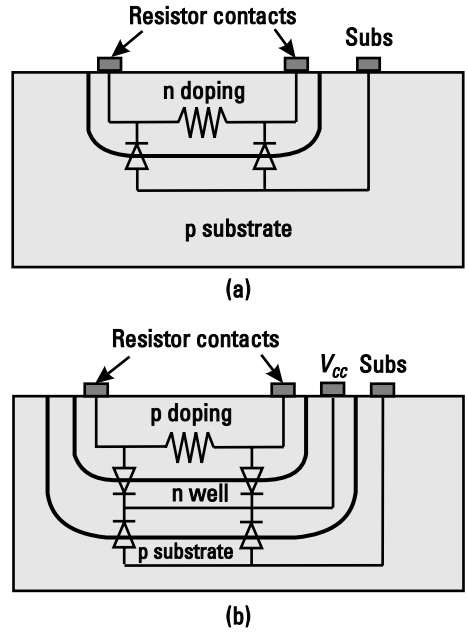

Figure 6.5 Diffusion resistors: (a) diffusion resistor without well isolation (n doping), and (b) diffusion resistor with well isolation (p doping).

Diffusion resistors are made by doping a layer of silicon to give it the desired resistivity, typically 1 kΩ/□ or more, and can be made with either p doping or n doping as shown in Figure 6.5. If n doping is used, then the structure can be quite simple because the edge of the doping region will form a pn junction with the substrate. Since this junction can never be forward-biased, current will not leak into the substrate. If, however, p doping is used, then it must be placed in an n-well to provide isolation from the substrate.

Note that all resistors regardless of what material they are made of will have a maximum current handling capability. This will usually be specified as a maximum allowable current per unit width. Thus, if resistors need to handle large currents, they must often be made wider than minimum width. The desired resistance is then achieved by adjusting the length of the device.

6.8 Metal-Insulator-Metal Capacitors and Stacked Metal Capacitors

We have already discussed that metal lines have parasitic capacitance associated with them. However, since it is generally desirable to make capacitance between metal as small as possible, they make poor deliberate capacitors. In order to improve this and conserve chip area, when capacitance between two metal lines is deliberate, the oxide between the two lines is thinned to increase the capacitance per unit area. This type of capacitor is called ametal-insulator-metal capacitor (MIM cap). More capacitance per unit area saves chip space. The capacitance between any two parallel-plate capacitors is given by (6.5) as discussed previously. Since

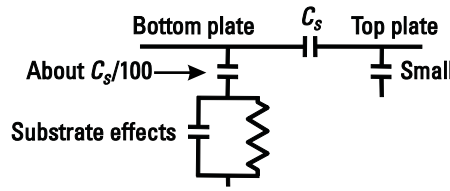

Figure 6.6 Model for an integrated MIM capacitor.

this expression holds for a wide range of applied voltages, these types of capacitors are extremely linear. However, if there is too much buildup of charge between the plates, the insulator can break down and conduct. This is of particular concern during processing of wafers, thus there are often rules (called antenna rules) governing how much metal can be connected directly to the capacitors. Typically in a modern process, protection devices are needed. A simple model for an integrated capacitor is shown in Figure 6.6.

Capacitors can also be made from complex configurations of metal. Modern processes often contain many layers of metal with very fine line width and spacing capability. Making an array of tall "posts" or "walls" by connecting multiple layers with vias is another way to make dense high-quality vertical capacitors. The more metal layers that are used, the more dense the capacitance per unit area; however, the closer the capacitor comes to the substrate, the more parasitics there are to deal with.

6.9 Applications of On-Chip Spiral Inductors and Transformers

The use of the inductor is illustrated in Figure 6.7, in which three inductors are shown in a circuit that is connected to a supply of value V_{CC}. A similar circuit that employs a transformer is shown in Figure 6.8. These two circuits are examples of LNAs, which will be discussed in detail in Chapter 7. The first job of the inductor is to resonate with any parasitic capacitance, potentially allowing higher frequency operation. A side effect (often wanted) is that such resonance results in filtering.

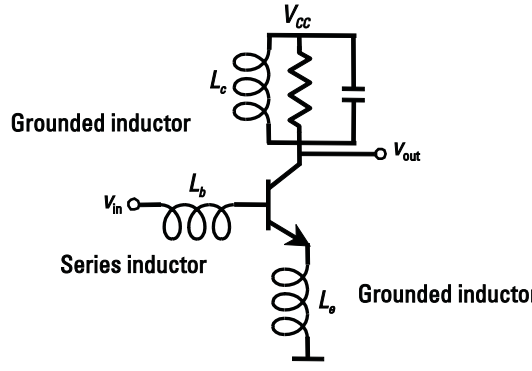

Figure 6.7 Application of inductors and capacitors.

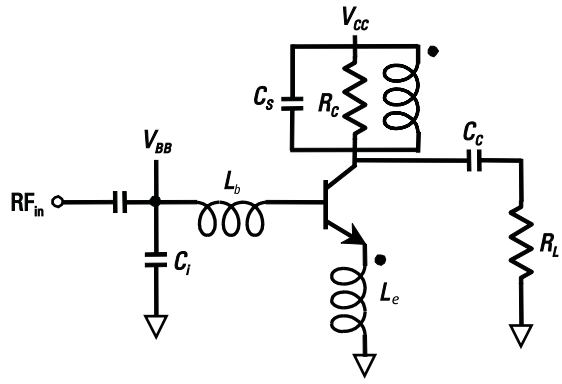

Figure 6.8 Application of a transformer.

Inductors L_b and L_e form the input match and degeneration while L_c forms a tuned load. As a load, or as emitter degeneration, one side of the inductor sees ac ground. This allows increased output swing, since there is ideally no dc voltage drop across the inductor. Similarly, the input series inductor has no dc voltage drop across it. The disadvantage is that, being in series, it has parasitic capacitance from both sides to the substrate. As a result, a signal can be injected into the substrate, with implications for noise and matching.

The transformer-based circuit as shown in Figure 6.8 and described by [4] has similar advantages to those described above for inductor-based tuned circuits. One difference is that the gain is determined (partially at least) by the transformer turns ratio, thus removing or minimizing dependence on transistor parameters. This has advantages since, unlike the transistor, the transformer has high linearity and low noise.

On the negative side, fully integrated transformers are lossy and more dif cult to model, and more sophisticated models are required.

6.10 Design of Inductors and Transformers

Special care must be taken to realize high-quality inductors and transformers or baluns monolithically. In silicon, they suffer from the presence of lossy substrates and nite metal resistance. However, over the past few years much research has been done in efforts to improve fabrication methods for building inductors, as well as modeling so that better geometries could be used in their fabrication.

When inductors are made in silicon technology with aluminum or copper interconnects, they suffer from the presence of relatively high-resistance interconnect structures and lossy substrates, typically limiting the Q to about 15 at around 2 GHz or about 18 to 25 at 5 GHz. This causes many high-speed RF components, such as voltage-controlled oscillators (VCOs) or power ampli ers using on-chip inductors to have limited performance compared to designs using off-chip components. The use of off-chip components adds complexity and cost to the design of these circuits, which has led to intense research aimed at improving the performance of on-chip inductors [5–19].

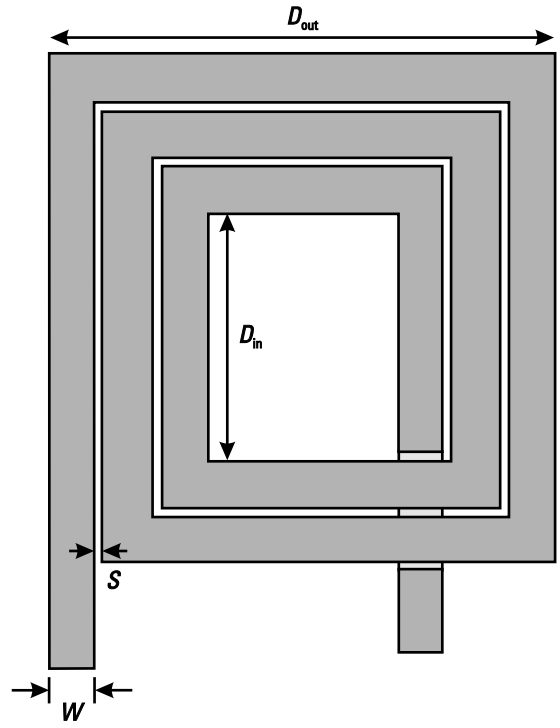

Figure 6.9 A conventional single-ended inductor layout.

Traditionally, due to limitations in modeling and simulation tools, inductors were made as square spirals as shown in Figure 6.9. The wrapping of the metal lines allows the ux from each turn to add, thus increasing the inductance per unit length of the structure, and providing a compact way of achieving useful values of inductance. Square inductors however, have less than optimum performance due to the ninety-degree bends present in the layout that add to the resistance of the structures. A better structure is shown in Figure 6.10 [7, 18]. Since this inductor is made circular, it has less series resistance. This geometry is more symmetric than

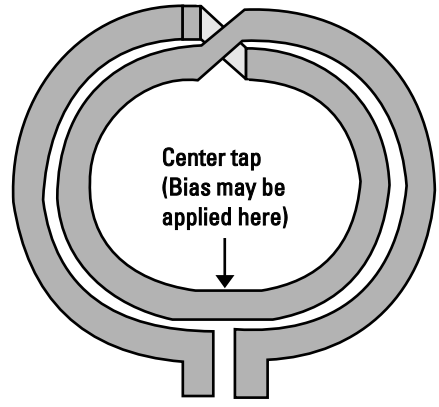

Figure 6.10 A circular differential inductor layout.

traditional inductors (its S-parameters look the same from either side). Thus, it can be used in differential circuits without needing two inductors to get good symmetry. Also, bias can be applied through the axis of symmetry of this structure if needed in a differential application (i.e., it is a virtual ground point).

6.11 Some Basic Lumped Models for Inductors

When describing on-chip inductors, it is useful to build an equivalent model for the structure. Figure 6.11 shows capacitance between lines, capacitance through the oxide, the inductance of the traces, series resistance and substrate effects. These effects are translated into the circuit model shown in Figure 6.12, which shows a number of nonideal components. R_S models the series resistance of the metal lines used to form the inductor. Note that the value of R_S will increase at higher frequencies due to the skin effect. C_{oxide} models the capacitance from the lines to the substrate. This is essentially a parallel plate capacitor formed between the inductor metal and the substrate. C_{sub} and R_{sub} model the losses due to magnetic effects, capacitance, and the conductance of the substrate. They are proportional to the area of the metal in the inductor, and their exact value depends on the properties of the substrate in question. C_{IW} models the interwinding capacitance between the traces. This is another parallel plate capacitor formed by adjacent metal lines. Note that in a regular inductor both sides are not symmetric, partly due to the added capacitance on one side of the structure caused by the underpass. The underpass connects the metal at the center of the planar coil with metal at the periphery.

The model for the symmetric or differential inductor is shown in Figure 6.13 [19]. Here the model is broken into two parts with a pin at the axis of symmetry where a bias can be applied if desired. Note also that since the two halves of the spiral are interleaved, there is magnetic coupling between both halves of the device. This is modeled by the coupling coefficient k. Note also that C_{oxide2}, R_{sub2} and C_{sub2} are connected to a virtual ground and therefore no longer have any effect on the RF performance of the device. More complex models for inductors at very high frequencies can include skin effect and image current in the substrate.

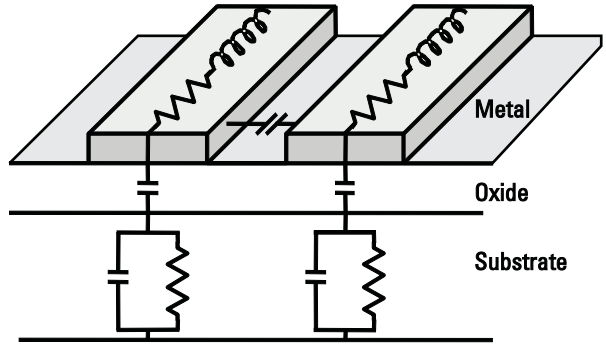

Figure 6.11 Elements used to build an inductor model.

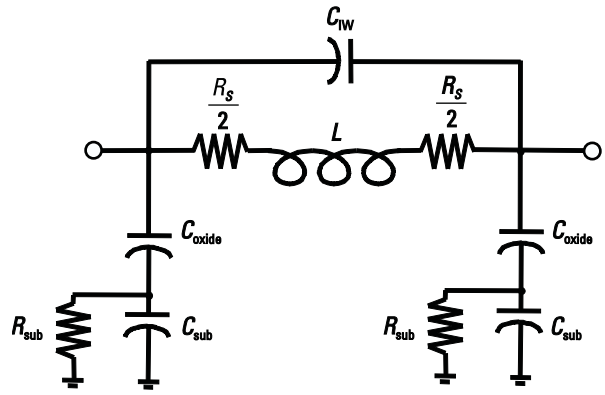

Figure 6.12 Basic model for a regular inductor.

6.12 Calculating the Inductance of Spirals

Formulas for calculating on-chip inductors have been proposed for both square and octagonal geometries [5]. The following simple expressions can be used:

$$L = 2.34 \; _{o} \frac{n^2 d_{avg}}{1 + 2.75} \tag{6.9}$$

for square inductors, wheren is the number of turns, d_{avg} is given by (see Figure 6.9)

$$d_{avg} = \frac{1}{2}(D_{out} + D_{in}) \tag{6.10}$$

and is given by:

$$= \frac{(D_{out} \quad D_{in})}{(D_{out} + D_{in})} \tag{6.11}$$

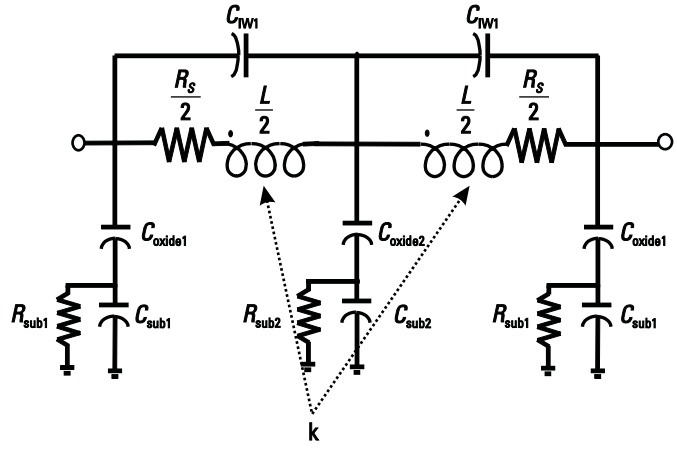

Figure 6.13 Basic model for a differential inductor.

and for octagonal inductors:

$$L = 2.25\ _{o}\frac{n^2 d_{avg}}{1+3.55} \tag{6.12}$$

The formulas can be quite accurate and their use will be demonstrated in Example 6.5. However, often it is easier to use simulators like ASITIC or other three-dimensional EM solvers. Since the substrate complicates matters, an EM simulator is often the only option for very complicated geometries. These can be quite slow, which makes them cumbersome to use as a design tool, but the speed is improving as computer power grows.

6.13 Self-Resonance of Inductors

At low frequencies, the inductance of an integrated inductor is relatively constant. However, as the frequency increases, the impedance of the parasitic capacitance elements starts to become signi cant. At some frequency, the admittance of the parasitic elements will cancel that of the inductor and the inductor will self-resonate. At this point, the imaginary part of the admittance will be zero. The inductance is nearly constant at frequencies much lower than the self-resonance frequency, however, as the self-resonance frequency is approached, the inductance rises and then abruptly falls to zero. Beyond the self-resonant frequency, the parasitic capacitance will dominate and the inductor will look capacitive. Thus, the inductor has a nite bandwidth over which it can be used. For reliable operation, it is necessary to stay well below the self-resonance frequency. Since parasitic capacitance increases in proportion to the size of the inductor, the self-resonant frequency decreases as the size of the inductor increases. Thus, the size of on-chip inductors that can be built is severely limited.

6.14 The Quality Factor of an Inductor

The quality factor, or Q, of a passive circuit element can be de ned as

$$Q = \frac{|\mathrm{Im}(Z_{ind})|}{|\mathrm{Re}(Z_{ind})|} \tag{6.13}$$

where Z_{ind} is the impedance of the inductor. This is not necessarily the most fundamental de nition of Q, but it is a good way to characterize the structure. A good way to think about this is that Q is a measure of the ratio of the desired quantity (inductive reactance) to the undesired quantity (resistance). Obviously a higher Q device is more ideal.

The Q of an on-chip inductor is affected by many things. At low frequencies, the Q tends to increase with frequency because the losses are relatively constant (mostly due to metal resistance R_s) while the imaginary part of the impedance is increasing linearly with frequency. However, as the frequency increases, currents

start to flow in the substrate through capacitive, and to a lesser degree, magnetic coupling. This loss of energy into the substrate causes an effective increase in the resistance. In addition, the skin effect starts to raise the resistance of the metal traces at higher frequencies. Thus, most integrated inductors have Qs that rise at low frequencies and then have some peak beyond which the losses make the resistance rise faster than the imaginary part of the impedance, and the Q starts to fall off again. Thus, it is easy to see the need for proper optimization to ensure that the inductor has peak performance at the frequency of interest.

Example 6.5: Calculating Model Values for the Inductor

Given a square inductor with the dimensions shown in Figure 6.14, determine a model for the structure including all model values. The inductor is made out of 3-μm-thick aluminum metal. The inductor is suspended over 5 μm of oxide above a substrate. The underpass is 1 μm Al and is 3 μm above the substrate. Assume the vias are lossless.

Solution:
We can start by estimating the inductance of the structure by using the formulas in Section 6.11 and referring to Figure 6.9.

$$d_{avg} = \frac{1}{2}(D_{out} + D_{in}) = \frac{1}{2}(270\ \mu m + 161\ \mu m) = 215.5\ \mu m$$

$$= \frac{(D_{out} \quad D_{in})}{(D_{out} + D_{in})} = \frac{(270 \quad m \quad 161 \quad m)}{(270 \quad m + 161 \quad m)} = 0.253$$

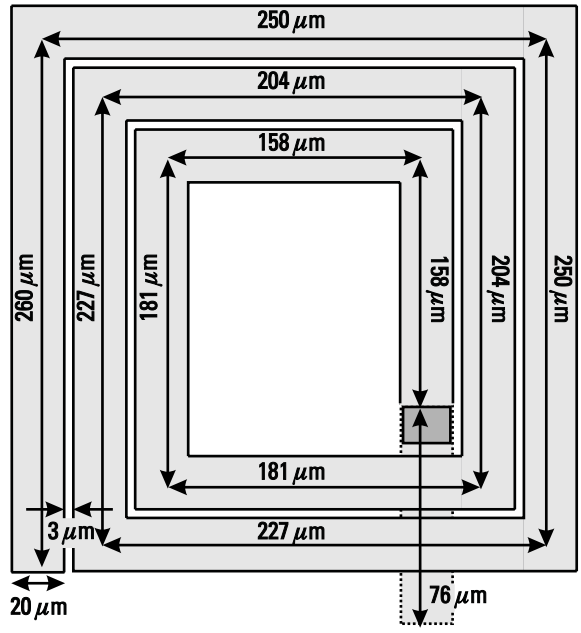

Figure 6.14 Inductor with dimensions.

$$L = 2.34\ _{o}\frac{n^2 d_{avg}}{1+2.75} = 2.34\times 4\ \times 10^{-7}\frac{N}{A^2}\frac{3^2\times 215.5\ \text{m}}{1+2.75\times 0.253} = 3.36\ \text{nH}$$

Next, let us estimate the oxide capacitance. First, the total length of the inductor metal is 2.3 mm. Thus, using (6.5) the formula for a parallel plate capacitor, the total capacitance through the oxide is:

$$C_{oxide} = \frac{_{o}\ _{r}A}{d} = \frac{8.85\times 10^{-12}\frac{C^2}{N\times m^2}\times 3.9\times 2.3\ \text{mm}\times 20\ \ \text{m}}{5\ \ \text{m}} = 317.6\ \text{fF}$$

The underpass must be taken into account here as well:

$$C_{underpass} = \frac{_{o}\ _{r}A}{d} = \frac{8.85\times 10^{-12}\frac{C^2}{N\times m^2}\times 3.9\times 76\ \ \text{m}\times 20\ \ \text{m}}{3\ \ \text{m}} = 17.4\ \text{fF}$$

Now we must consider the interwinding capacitance, noting that this length is shorter than the total length of about 2.3 mm by nearly the length of the inner turn, or about 0.62 mm:

$$C_{IW} = \frac{_{o}\ _{r}A}{d} = \frac{8.85\times 10^{-12}\frac{C^2}{N\times m^2}\times 3.9\times (2.3\ \text{mm}\ \ 0.62\ \text{mm})\times 3\ \ \text{m}}{3\ \ \text{m}} = 58\ \text{fF}$$

The dc resistance of the line can be calculated from (6.2):

$$R_{DC} = \frac{L}{WH} = \frac{3\ \ \text{m}\ \ \text{cm}\times 2.3\ \text{mm}}{20\ \ \text{m}\times 3\ \ \text{m}} = 1.15$$

The skin effect will begin to become important when the thickness of the metal is two skin depths. Using (6.4), this will happen at a frequency of:

$$f = \frac{}{^{2}} = \frac{3\ \ \text{cm}}{\times 4\ \ \times 10^{-7}\times (1.5\ \ \text{m})^2} = 3.38\ \text{GHz}$$

Let us ignore the resistance in the underpass. Thus, above 3.38 GHz the resistance of the line will be a function of frequency. We will now modify the area used in (6.2) to take into account the skin depth effect:

$$R_{ac}(f) = \frac{L}{Wt\quad W\quad 2\sqrt{\frac{}{f}}\div t\quad 2\sqrt{\frac{}{f}}\div}$$

The other thing that must be considered is the substrate. This is an issue for which we really do need a simulator. However, as mentioned above, the capacitance and resistance will be a function of the area. This also means that once several structures in a given technology have been measured, it may be possible to predict these values for future structures. For this example, assume that reasonable

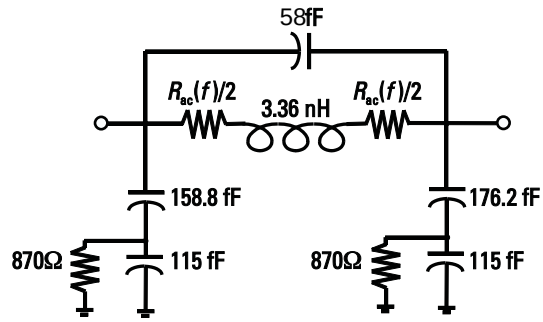

Figure 6.15 Inductor pi model with numbers. $R_{ac}(f)$ has a low frequency value of 1.15W.

values for the tting parameters have been determined. Thus, R_{sub} and C_{sub} could be something like 870W and 115 fF. The complete model with values is shown in Figure 6.15.

Note that parasitic capacitances and inductances obtained by these closed form equations are estimates that may be useful in circuit simulation and initial layout construction. However, a nal design should include EM simulation or measured veri cation of inductor model parameters.

Example 6.6: Determining Inductance, Q, and Self-Resonant Frequency

Take the model just created for the inductor in the previous example, and compute the equivalent inductance, and Q versus frequency. Also, nd the self-resonance frequency. Assume that the side of the inductor with the underpass is grounded.

Solution:

The equivalent circuit in this case is as shown in Figure 6.16.

This is just an elementary impedance network, so we will skip the details of the analysis and give the results. The inductance is computed by taking the imaginary part of Z_{in} and dividing by 2 f. The Q is computed as in (6.13), and the results are shown in Figure 6.17. The self-resonant frequency of this structure can be read from the graph. It is 7.8 GHz.

This example shows a peak Q of 21, but in reality due to higher substrate losses and line resistance leading up to the inductor, the Q will be somewhat lower than shown here, although in most respects this example has shown very realistic results.

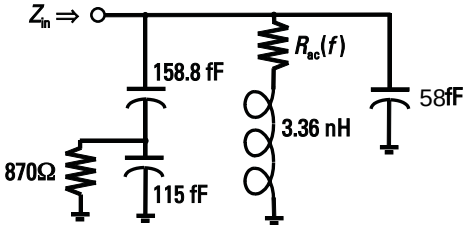

Figure 6.16 Inductor model with one side grounded.

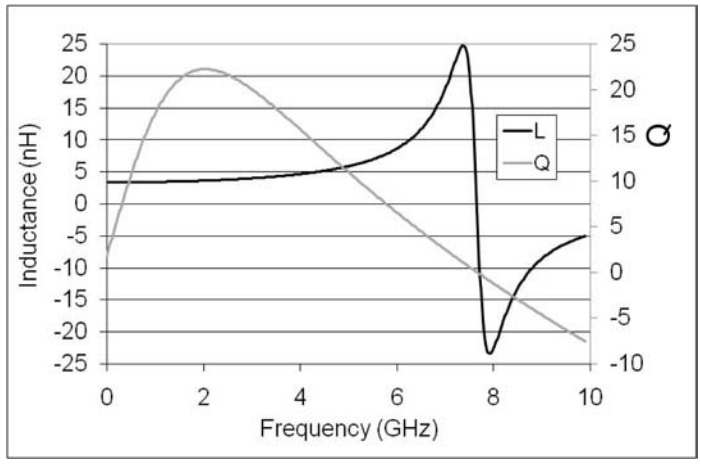

Figure 6.17 Inductor plot of L and Q versus frequency.

6.15 Characterization of an Inductor

Once some inductors have been built and measured, S parameter data will then be available for these structures. It is then necessary to take these numbers and convert them for example into inductance, Q, and self-resonance frequency.

The de nitions of Q have already been given in (6.13) and L is equal to the imaginary part of the impedance. These de nitions seem like simple ones, but the impedance still needs to be de ned. Traditionally, we have assumed that one port of the inductor is grounded. In such a case, we can de ne the impedance seen from port 1 to ground.

Starting with the Z parameter matrix (which can be easily derived from S parameter data):

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix} \tag{6.14}$$

Since the second port is grounded, V_2=0. Thus, two equations result:

$$\begin{aligned} V_1 &= Z_{11} I_1 + Z_{12} I_2 \\ 0 &= Z_{21} I_1 + Z_{22} I_2 \end{aligned} \tag{6.15}$$

The second equation can be solved for I_2:

$$I_2 = \frac{Z_{21} I_1}{Z_{22}} \tag{6.16}$$

Thus, I_2 can now be removed from the rst equation, and solving for $Z_{in} = Z_{Port1} = V_1/I_1$:

$$Z_{Port1} = Z_{11} \frac{Z_{12} Z_{21}}{Z_{22}} \tag{6.17}$$

Equivalently, if we look from port 2 to ground the impedance becomes:

$$Z_{Port2} = Z_{22} - \frac{Z_{12}Z_{21}}{Z_{11}} \tag{6.18}$$

Note that, referring to Figure 6.12, this effectively grounds out C_{oxide}, R_{sub}, and C_{sub} on one side of the inductor. Thus, the Q will not necessarily be the same looking from both ports. In fact, the Q will be marginally higher in the case of a regular structure looking from the side with no underpass, as there will be less loss. Also, note that the side with no underpass will have a higher self-resonance frequency.

Often designers want to use inductors in a differential con guration. This means that both ends of the inductor are connected to active points in the circuit and neither side is connected to ground. In this case, we can de ne the impedance seen between the two ports:

Starting again with the Z parameters, the voltage difference applied across the structure is now:

$$V_1 - V_2 = Z_{11}I_1 + Z_{12}I_2 - Z_{21}I_1 - Z_{22}I_2 \tag{6.19}$$

$$V_1 - V_2 = I_1(Z_{11} - Z_{21}) - I_2(Z_{22} - Z_{12}) \tag{6.20}$$

Because the structure is symmetric we make the assumption that $I_1 = -I_2$ thus:

$$Z_{Diff} = \frac{V_1 - V_2}{I_1} = Z_{11} + Z_{22} - Z_{12} - Z_{21} \tag{6.21}$$

In this case, the substrate capacitance and resistance from both halves of the inductor are in series. When the inductor is excited in this mode, it “sees” less loss and will give a higher Q. Thus, the differential Q is usually higher than the single-ended Q. The self-resonance of the inductor in this mode will also be higher than the self-resonance frequency looking from either side to ground. As well, the frequency at which the differential Q peaks is usually higher than for the single-ended excitation. Care must be taken, therefore, when optimizing an inductor for a given frequency, to keep in mind its intended con guration in the circuit.

It is important to note that every inductor has a differential Q and a single-ended Q regardless of its layout. Which Q should be used in analysis depends on how the inductor is used in a circuit.

6.16 Some Notes about the Proper Use of Inductors

Designers are very hesitant to place a nonsymmetric regular inductor across a differential circuit. Instead, two regular inductors are usually used. In this case, the center of the two inductors is effectively AC grounded and the effective Q for the two inductors is equal to their individual single-ended Qs. To illustrate this point, take a simpli ed model of an inductor with only substrate loss as shown in Figure 6.18. In this case the single-ended Q with one terminal grounded is given by:

$$Q_{SE} = \frac{R}{L} \tag{6.22}$$

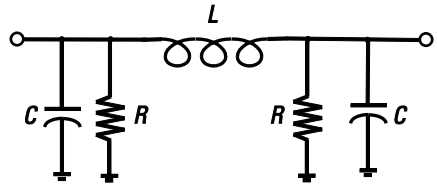

Figure 6.18 Simpli ed inductor model with only substrate loss.

and the differential Q is given by:

$$Q_{Diff} = \frac{2R}{L} \tag{6.23}$$

Note that in the differential case that the two resistors are in series for a total resistance of 2R.

Now if two inductors are placed in series as shown Figure 6.19, the differential Q of the overall structure is given by:

$$Q_{Diff2} = \frac{2R}{2L} = \frac{R}{L} = Q_{SE} \tag{6.24}$$

Thus, the differential Q of the two inductors is equal to the single-ended Q of one of the inductors. The advantage of using symmetric structures should be obvious. Note that, here, substrate losses have been assumed to dominate. If the series resistance is dominant in the structure, then using this con guration will be less advantageous. However, due to the mutual coupling of the structure, this con guration would still be preferred, as it makes more ef cient use of the chip area.

If using a differential inductor in the same circuit, the designer would probably use only one structure. In this case, the effective Q of the circuit will be equal to the differential Q of the inductor. Note that if a regular inductor were used in its place, the circuit would see its differential Q as well.

When using a regular inductor with one side connected to ground, the side with the underpass should be the side that is grounded, as this will result in a higher Q and a higher self-resonance frequency.

Example 6.7: Single-Ended Versus Differential Q

Take the inductor of Example 6.5 and compute the single-ended Q from both ports as well as the differential Q. Also, compare the self-resonant frequency under these three conditions.

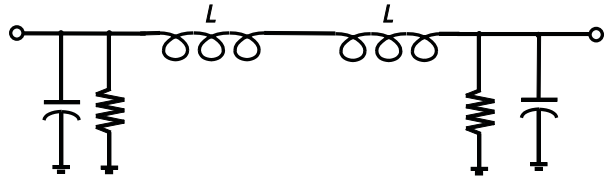

Figure 6.19 Two simpli ed inductors connected in series with only substrate loss.

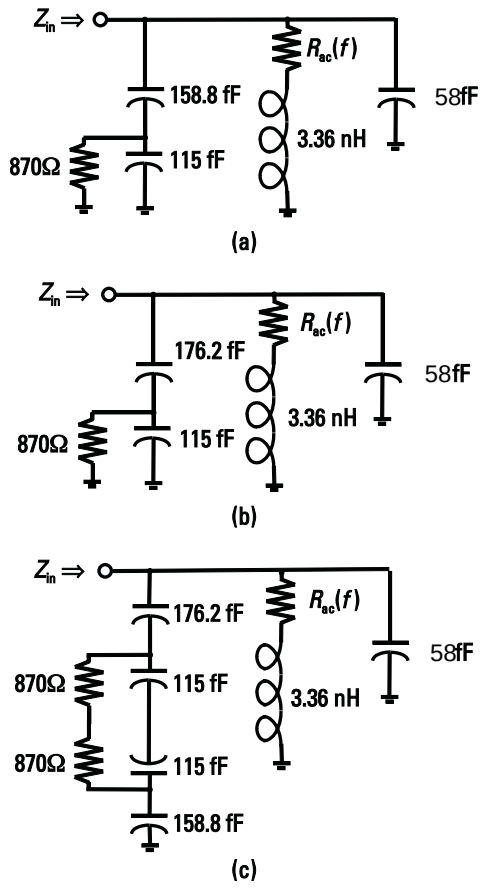

Figure 6.20 Equivalent circuits for the case where (a) the side with the underpass grounded, (b) where the underpass is not grounded, and (c) the circuit is driven differentially.

Solution:
As before, equivalent circuits can be made from the model in the previous example (see Figure 6.20). Note that the substrate connections from either side of the circuit appear in series when the device is operated differentially as shown in Figure 6.20(c).

It is a matter of elementary circuit analysis to compute the input impedance of these three networks. The inductance of all three is shown in Figure 6.21. Note that the case where the underpass is not grounded, the circuit has the lowest self-resonant frequency, while the differential con guration leads to the highest self-resonant frequency.

The Q of these three networks is plotted in Figure 6.22. Note here that at low frequencies where substrate effects are less important, the Qs are all equal, but as the frequencies increase, the case where the circuit is driven differentially is clearly better. In addition, in this case, theQ keeps rising to a higher frequency and higher overall value.

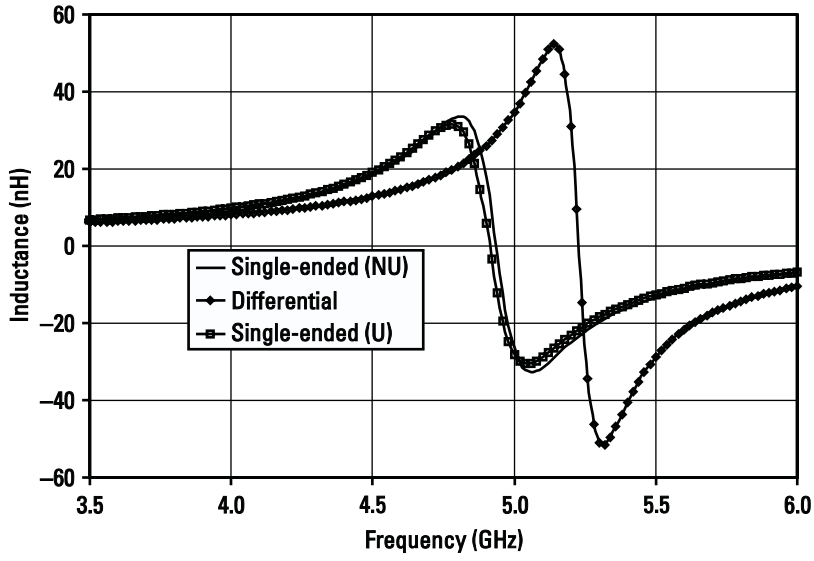

Figure 6.21 The inductance plotted versus frequency for the three modes of operation.

6.17 Layout of Spiral Inductors

The goal of any inductor layout is to design a spiral inductor of speci ed inductance, with Q optimized for best performance at the frequency of interest. In order to achieve this, careful layout of the structure is required. The resistance of the metal lines causes the inductor to have a high series resistance, limiting its performance at low frequencies, while the proximity of the substrate causes substrate loss, raising the effective resistance at higher frequencies. Large coupling between the inductor

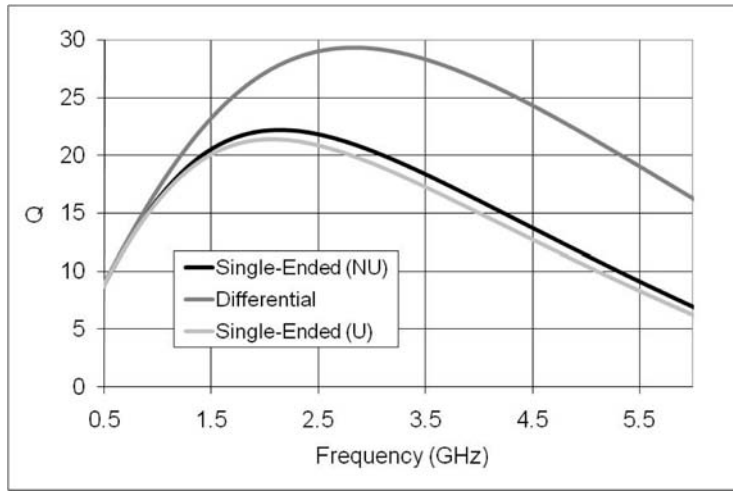

Figure 6.22 The Q plotted versus frequency for the three modes of operation.

and the substrate also causes the structures to have low self-resonance frequencies. As a result, there are limitations on the size of the device that can be built.

Traditionally, on-chip inductors have been square as shown in Figure 6.9. This is because these have been easier to model than geometries that are more complicated. A square geometry is by no means optimal, however. The presence of the 90° bends adds unnecessary resistance to the structure, and as the structure is made circular, the performance will improve.

Some guidelines for optimum layout will now be provided. These rules are based on considerations of the effect of geometry on the equivalent model shown in Figure 6.12.

1. Line spacing: At low frequencies (2 GHz or less) keep the line spacing as tight as possible. At higher frequencies, due to coupling between turns, larger spacing may be desirable.
2. Line width: Increasing metal width will reduce the inductance (fewer turns in a given area and also less inductance per unit length) and will decrease the series resistance of the lines at low frequencies. Large inductance area means bigger capacitance, which means lower self-resonance, and more coupling of current into the substrate. Therefore as W goes up, inductance comes down, and the frequency of Q_{peak} gets lower (and vice versa). Line widths for typical 1- to 5-nH inductors in the 2- to 5-GHz range would be expected to be from 5 to 25 μm.
3. Area: Bigger area means that more current is present in the substrate, so high frequency losses tend to be increased. Bigger area (for the same line width) means longer spirals, which means more inductance. Therefore as the area goes up, inductance goes up, and the frequency of Q_{peak} gets lower (and vice versa).
4. Number of turns: This is typically a third degree of freedom. While more turns will result in higher inductance, it will also increase interwinding capacitance and hence reduce the self-resonance frequency. For this reason, it is often better to use fewer turns and instead increase inductance by increasing outer diameter, provided that the inductor does not get to be huge. Huge is of course a relative term and it ultimately is up to the designer to decide how much space they are willing to devote to the inductor layout. Inner turns add less to the inductance but more resistance, thus it is best to keep the inductor hollow. By changing the area and the line width, the peak frequency and inductance can be ne-tuned.

6.18 Isolating the Inductor

Inductors tend to be extremely large structures, and as such they tend to couple signals into the substrate; therefore isolation must be provided. Typically, a ring of substrate contacts is added around each inductor. These substrate contacts are usually placed at a distance of about ve line widths away from the inductor. The presence of a patterned (slotted) ground shield, discussed in the next section, may also help in isolating the inductor from the substrate.

6.19 The Use of Slotted Ground Shields and Inductors

In an inductor, currents flow into the substrate through capacitive coupling and are induced in the substrate through magnetic coupling. Current flowing in the substrate causes additional loss. Of the two, generally capacitive coupling is the more dominant loss mechanism. One method to reduce substrate loss is to place a ground plane above the substrate, preventing currents from entering the substrate [12]. However, with a ground plane, magnetically generated currents will be increased, reducing the inductance. One way to get around this problem is to pattern the ground plane such that magnetically generated currents are blocked from flowing. An example of a patterned ground shield designed for a square inductor is shown in Figure 6.23. Slots are cut into the plane perpendicular to the direction of magnetic current flow. The ground shield has the disadvantage of increasing capacitance to the inductor, causing its self-resonant frequency to drop significantly. For best performance, the ground shield should be placed far away from the inductor, but remain above the substrate. In a typical bipolar process, the polysilicon layer is a good choice.

The model for the ground-shielded inductor compared to the standard inductor model is shown in Figure 6.24. For the ground-shielded inductor, the lossy substrate capacitance has been removed leaving only the lossless oxide capacitance and the parasitic resistance of the shield. As a result, the inductor will have a higher Q.

6.20 Basic Transformer Layouts in IC Technologies

Transformers in silicon are more complicated than inductors and therefore somewhat harder to model. Transformers or baluns consist of two interwound spirals that are magnetically coupled. An example layout of a basic structure is shown Figure 6.25. In this figure, two spirals are interwound in a 3:3 turns ratio structure. The structure can be characterized by a primary and secondary inductance and a mutual inductance or coupling factor, which describes how efficiently energy can be transferred from one spiral to the other. A symmetric structure with a turns ratio of 2:1 is shown in Figure 6.26.

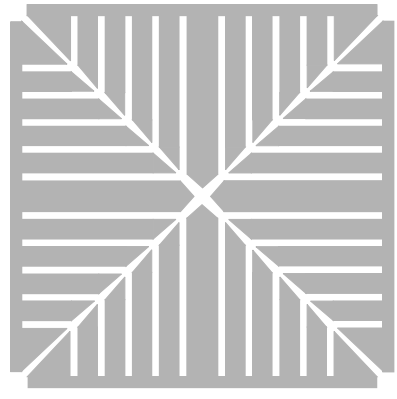

Figure 6.23 Patterned ground shield for a square spiral inductor (inductor not shown).

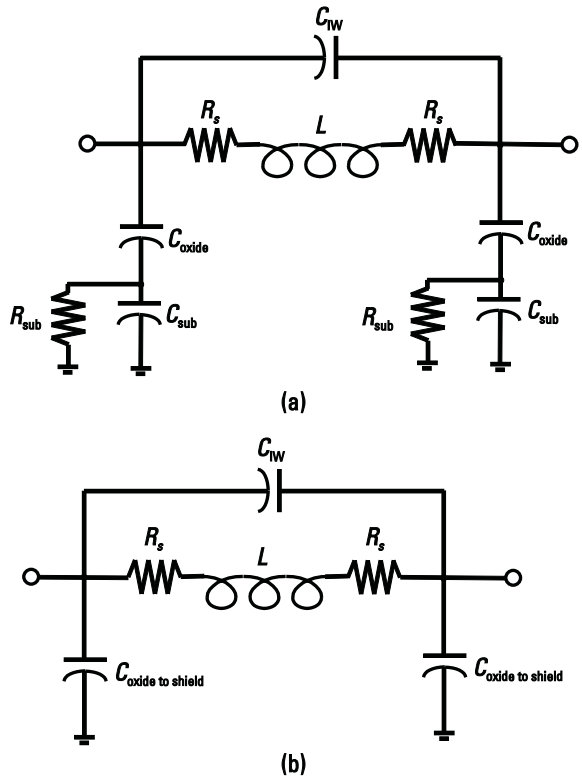

Figure 6.24 Comparison of the basic model for: (a) regular and (b) shielded inductors.

A simpli ed transformer model is shown in Figure 6.27. This is modeled as two inductors, but with the addition of coupling coef cient k between them, and interwinding capacitance C_{IW} from input to output.

Example 6.8: Placing the Dots

Place the dots on the transformer shown in Figure 6.26.

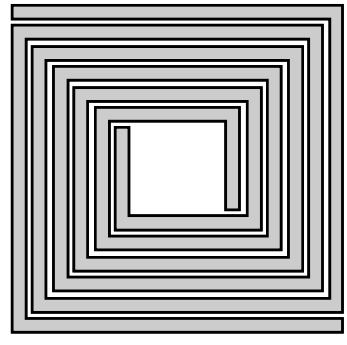

Figure 6.25 Sample layout of two interwound inductors forming a transformer.

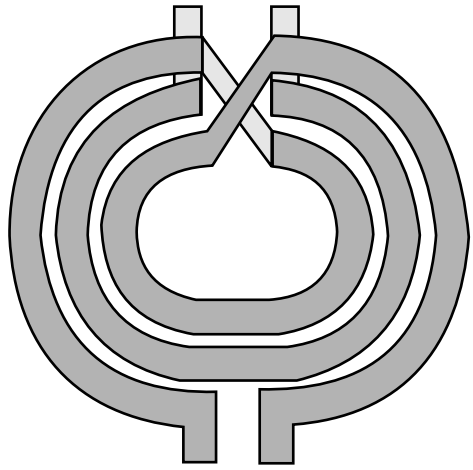

Figure 6.26 Sample layout of a circular 2:1 turns ratio transformer.

Solution:
We will start by assuming that a current is owing in the primary winding, as shown in Figure 6.28(a). This will cause a ux to ow into the page at the center of the winding, as shown in Figure 6.28(b). Thus, in order for a current in the secondary to reinforce the ux, it must ow in the direction shown in Figure 6.28(c). Therefore, the dots go next to the ports where the current ows out of the transformer, as shown in Figure 6.28(d).

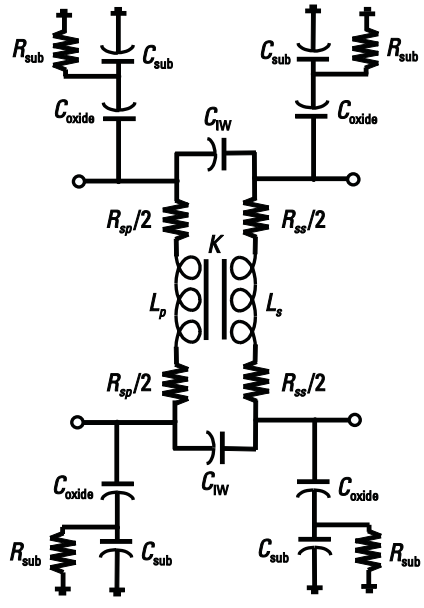

Figure 6.27 Basic model of a transformer.

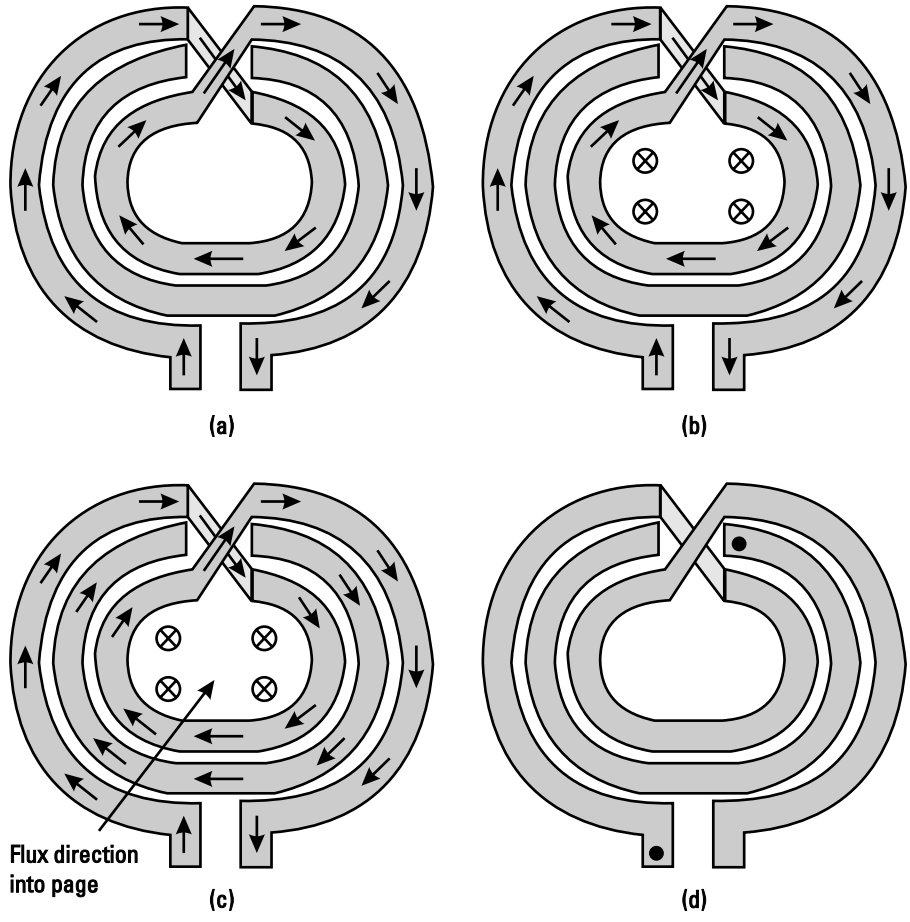

Figure 6.28 Determining dot placement: (a) arbitrary current ow, (b) direction of ux, (c) secondary current ow that adds to the ux, and (d) dot placement.

6.21 Multilevel Inductors

Inductors can also be made using more than one level of metal. Especially in modern processes, which can have as many as ve or more metal layers, it can be advantageous to do so. There are two common ways to build multilevel inductors. The rst is simply to strap two layers or more layers of metal together with vias to decrease the effective series resistance. This will increase the Q but at the expense of increased capacitance to the substrate and a resultant decrease in self-resonant frequency. This technique is of bene t for small inductors for which the substrate loss is not dominant and that are at low enough frequency, safely away from the self-resonant frequency.

The second method is to connect two or more layers in series. This results in increased inductance for the same area, or allows the same inductance to be realized in a smaller area. A drawing of a two-level inductor is shown in Figure 6.29. Note that the uxes through the two windings will reinforce one another and the total inductance of the structure will be $L_{top} + L_{bottom} + 2M$ in this case. If perfect

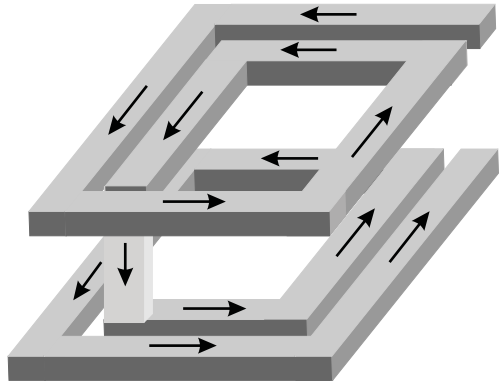

Figure 6.29 Three-dimensional drawing of a multilevel inductor.

coupling is assumed, and the inductors are of equal size, then this gives $L_1 n^2$. In general, this is a factor of n^2 more inductance, where n is the number of levels. Thus, this is a way to get larger inductance without using as much chip area.

To estimate the inductance of a multilevel inductor, the formulas presented in Section 6.11 can be used on the individual layers and the result is then multiplied by n^2. Note that actual inductance will be smaller because of imperfect coupling.

To determine the capacitance associated with the inductor, we consider the top and bottom spirals as two plates of a capacitor with total distributed capacitance C_1 [17]. In addition, we consider the bottom spiral and the substrate to form a distributed capacitance C_2. Now the total equivalent capacitance of the structure can be approximated. First, note that if a voltage V_1 is applied across the terminals of the inductor, the voltage across C_1 will go from V_1 at the terminals down to zero at the via. Similarly, the voltage across C_2 will go from zero at the terminal (assuming this point is grounded) to $V_1/2$ at the via.

Thus, the total energy stored in C_1 is

$$E_{C_1} = \frac{1}{2} C_1 V_1^2 \int_0^1 (1 - x)^2 dx = \frac{1}{2} \frac{C_1}{3} V_1^2 \tag{6.25}$$

where x is a dummy variable representing the normalized length of the spiral. Note that $V_1(1 - x)$ is an approximation to the voltage across C_1 at any point along the spiral. Thus the equivalent capacitance of C_1 is

$$C_{eq1} = \frac{C_1}{3} \tag{6.26}$$

The total capacitance can be found in C_2 in much the same way.

$$E_{C_2} = \frac{1}{2} C_2 \left(\frac{V_1}{2} \right)^2 \int_0^1 x^2 dx = \frac{1}{2} \frac{C_2}{3 \times 2^2} V_1^2 \tag{6.27}$$

Note that $V_1/2 \times x$ approximates the voltage across C_2 at any point along the spiral.

$$C_{eq2} = \frac{C_1}{3 \times 2^2} \tag{6.28}$$

Thus, the total capacitance is:

$$C_{eq} = \frac{C_1}{3} + \frac{C_2}{12} \tag{6.29}$$

Note that the capacitor C_2 is of less importance than C_1. Thus, it would be advantageous to space the two spirals far apart even if this means there is more substrate capacitance (C_2). Note also that C_1 will have a low loss associated with it, since it is not dissipating energy in the substrate.

6.22 Characterizing Transformers for Use in ICs

Traditionally transformers are characterized by their S-parameters. While correct, this gives little directly applicable information about how the transformer will behave in an application when loaded with impedances other than 50 Ω. It would be more useful to extract an inductance and Q for both windings and plot the coupling (k factor) or mutual inductance for the structure instead. These properties have the advantage that they do not depend on the system reference impedance.

In the following narrowband model, all the losses are grouped into a primary and secondary resistance, as shown in Figure 6.30.

The model parameters can be found from the Z-parameters starting with

$$\left.\frac{V_1}{I_1}\right|_{I_2=0} = Z_{11} = R_p + j\omega(L_p - M) + j\omega M = R_p + j\omega L_p \tag{6.30}$$

Similarly:

$$Z_{22} = R_s + j\omega L_s \tag{6.31}$$

Thus, the inductance of the primary and secondary and the primary and secondary Q can be defined as

$$L_s = \frac{\mathrm{Im}(Z_{22})}{\omega} \tag{6.32}$$

$$L_p = \frac{\mathrm{Im}(Z_{11})}{\omega} \tag{6.33}$$

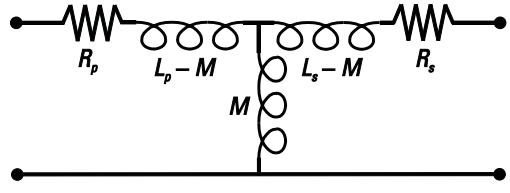

Figure 6.30 Narrowband equivalent model for a transformer.

$$Q_s = \frac{\mathrm{Im}(Z_{22})}{\mathrm{Re}(Z_{22})} \tag{6.34}$$

$$Q_p = \frac{\mathrm{Im}(Z_{11})}{\mathrm{Re}(Z_{11})} \tag{6.35}$$

The mutual inductance can also be extracted as:

$$\left.\frac{V_1}{I_2}\right|_{I_1=0} = Z_{12} = j\ M \tag{6.36}$$

Therefore,

$$M = \frac{Z_{12}}{j} = \frac{Z_{21}}{j} \tag{6.37}$$

6.23 On-Chip Transmission Lines

Any on-chip interconnect can be modeled as a transmission line. Transmission line effects on-chip can often be ignored if transmission lines are signi cantly shorter than a quarter wavelength at the frequency of interest. Thus, transmission line effects are often ignored for frequencies between 0 and 5 GHz. However, as higher frequency applications become popular, these effects will become more important.

One of the simplest ways to build a transmission line is by placing a conductor near a ground plane separated by a dielectric, as shown in Figure 6.31. Another type of transmission line is acoplanar waveguide, as shown in Figure 6.32. Note that in this case a ground plane is not needed, although it will be present in an IC.

The effect of on-chip transmission lines is to cause phase shift and possibly some loss. Since dimensions in an IC are typically much less than on the printed circuit board (PCB), this is often ignored. The magnitude of these effects can be estimated with a simulator (for example, Agilent's LineCalc), and included as transmission lines in the simulator if it turns out to be important. As a quick estimate for delay, consider that in free space (vacuum) a 1-GHz signal has a wavelength of 30 cm. However, oxide has $_{ox}$ = 3.9, which slows the speed of propagation by $\sqrt{3.9} = 1.975$ 2. Thus, the wavelength is 15.19 cm and the resultant phase shift is about 2.37°/mm/GHz. The on-chip line can be designed to be a nearly lossless

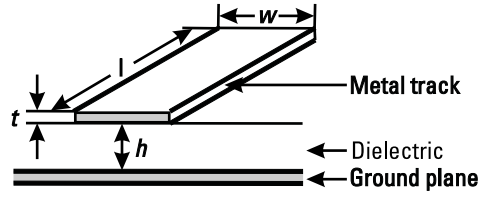

Figure 6.31 Microstrip transmission line.

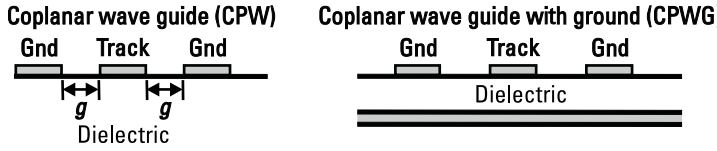

Figure 6.32 Coplanar wave guide transmission lines.

transmission line by including a shield metal underneath. Without such a shield metal, the substrate forms the ground plane and there will be losses. On silicon, a 50W line is about twice as wide as the dielectric thickness. Some quick simulations show that for a 4-µm dielectric, widths of 3, 6, and 12 µm result in about 72, 54, and 36W. Characteristic impedance can be calculated by the formula

$$Z_o = \sqrt{\frac{L}{C}} \tag{6.38}$$

where L and C are the per-unit inductance and capacitance, respectively. We note that if a length of transmission line is necessary on-chip, it may be advantageous to design it with a characteristic impedance higher than 50W, as this will result in less current necessary in the circuits matched to it.

6.23.1 Effect of Transmission Line

Matching an ampli er must include the effect of the transmission line up to the matching components. This transmission line causes phase shift (seen as rotation around the center of the Smith chart). If this effect is not considered, matching components can be completely incorrect. As an example, consider an RF circuit on a printed circuit board with off-chip impedance matching. At 5 GHz, with a dielectric constant of 4, a quarter wavelength is about 7.5 mm. This could easily be the distance to the matching components, in which case the circuit impedance has been rotated halfway around the Smith chart and impedance matching calculated without taking this transmission line into account would result in completely incorrect matching. For example, if a parallel capacitor is needed directly at the RF circuit, at a quarter wavelength distant a series inductor will be needed.

A number of tools are available that can do calculations of transmission lines and simulators can directly include transmission line models to see the effect of these lines.

6.23.2 Transmission Line Examples

At RF frequencies, any track on a printed circuit board behaves as a transmission line, such as a microstrip line (MLIN) (Figure 6.31) a coplanar waveguide (CPW) line (Figure 6.32), or a coplanar wave guide with ground (CPWG). Differential lines are often designed as coupled microstrip lines (MCLINs) (Figure 6.33), or they can become coupled simply because they are close together, for example at the pins of an integrated circuit. For these lines, differential and common mode impedance can be de ned (in microwave terms, these are described as odd-order and even-order

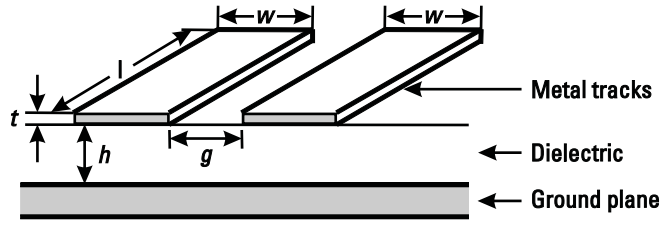

Figure 6.33 Coupled microstrip lines.

impedance, respectively). On an integrated circuit, all lines are transmission lines, even though it may be possible to ignore transmission line effects for short lines. The quality of such transmission lines may suffer due to the lossy ground plane (the substrate) or because of a poor connection between coplanar ground and substrate ground.

For example, a 400-µm by 3-µm line on-chip, with oxide (dielectric) thickness of 4 µm is simulated to have a characteristic impedance of about 70W. The capacitance is estimated by:

$$C = \frac{3.9 \times 8.85 \times 10^{-12}\,\text{F/m} \times 3 \times 10^{-6} \times 400 \times 10^{-6}}{4 \times 10^{-6}} = 10.35\,\text{fF}$$

Because of fringing and edge effects, the capacitance is probably more like 20 fF with the result that the inductance L is about 0.104 nH. Note for a wider line, such as 6µm, the capacitance is estimated to be about 30 fF, the characteristic impedance is closer to 50W, and inductance is about 0.087 nH.

In making use of a simulator to determine transmission line parameters, one needs to specify the dielectric thickness, line widths and gaps, dielectric constant, loss tangent, metal conductor conductivity, and thickness. Most simulators need dimensions speci ed in mils (thousandths of an inch) where a mil is equal to 25.4 µm. A typical dielectric thickness for a double-sided printed circuit board is 40 to 64 mil. Multilayered boards can have effective layers that are 10 mil or even less. Surface material is often copper with a thickness typically speci ed by weight, for example, half-ounce copper translates to 0.7 mil. Table 6.4 shows parameters for a variety of materials, including on-chip material (SiO_2, Si, GaAs) printed circuit board material (FR4, 5880, 6010) and some traditional dielectric material for microwave (e.g., ceramic). In simulators, the conductivity is typically speci ed relative to the conductivity of gold. Thus, using Table 6.4, $_{Au} = 1.42\ _{Cu}$, or $_{Cu} = 0.70\ _{Au}$.

Table 6.4 Properties of Various Materials

Material	Loss Tangent	Permittivity	Material	Loss Tangent	Permittivity
SiO	0.004–0.04	3.9	Al_2O_3(Ceramic)	0.0001	9.8
Si	0.015	11.9	Sapphire	0.0001	9.4
GaAs	0.002	12.9	Quartz	0.0001	3.78
FR4	0.022	4.3	6010	0.002	10.2
5880	0.001	2.20			

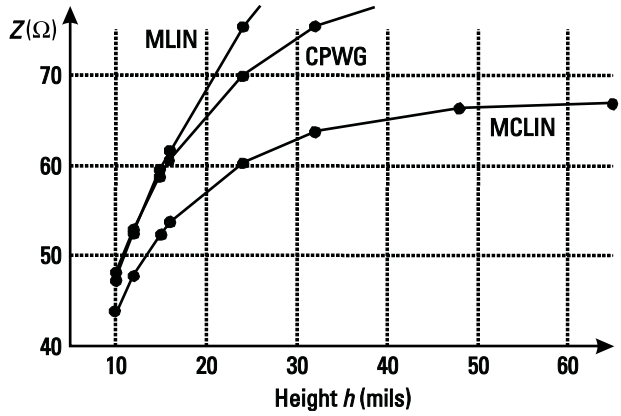

Figure 6.34 Impedance versus dielectric thickness for FR4 with line width of 20 mils at 1.9 GHz.

Example 6.9: Calculation of Transmission Lines

Using a simulator, determine line impedance at 1.9 GHz versus dielectric thickness for microstrip lines, coupled microstrip lines, and coplanar waveguide with a ground plane. Use FR4 material with a dielectric constant of 4.3, and 0.7-mil copper, a line width of 20 mil, and a 20-mil gap or space between the lines.

Solution:
Calculations were done and the results are shown in Figure 6.34. It can be seen that 50Wis realized with a dielectric thickness of about 11 mil for the microstrip line and the coplanar waveguide with ground and about 14 mil for the coupled microstrip lines. Thus, the height is just over half of the linewidth. It can also be seen that a

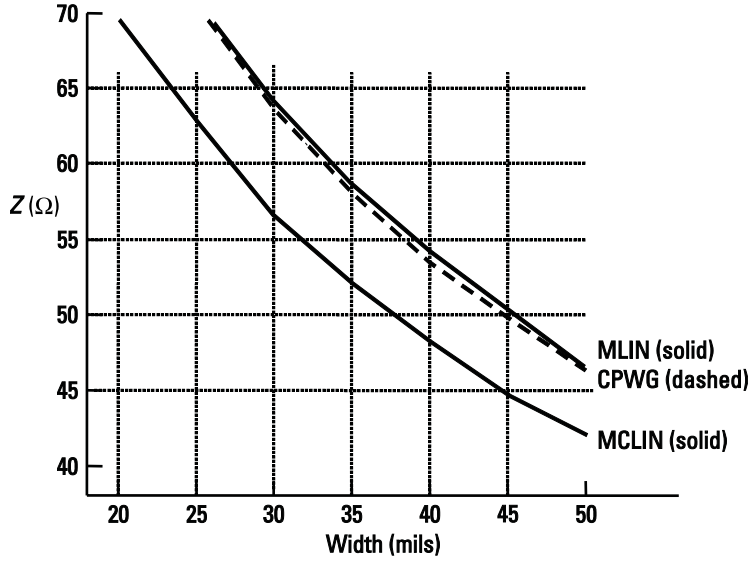

Figure 6.35 Impedance versus track width with dielectric thickness of 15 mils, gap or space of 20 mils, and dielectric constant of 2.2 at 1.9 GHz.

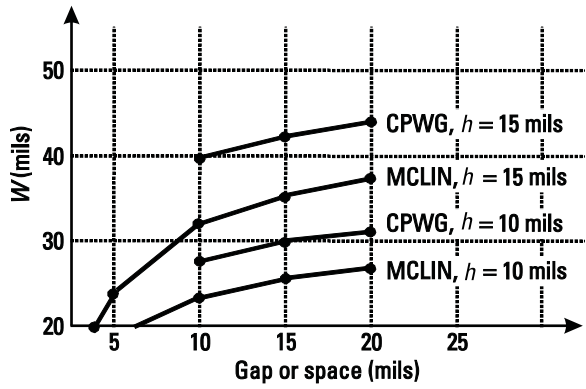

Figure 6.36 Track width versus gap or space to result in Z = 50W for coupled microstrip lines and coplanar waveguide.

microstrip line and a coplanar waveguide with ground have very similar behavior until the dielectric height is comparable to the gap dimension.

Example 6.10: Transmission Lines

Using a simulator, determine line impedance at 1.9 GHz versus line width, gap, and space for microstrip lines, coupled microstrip lines, and coplanar waveguide with a ground plane. Use material with a dielectric constant of 2.2 and height of 15 mils, and 0.7-mil copper.

Solution:
Calculations were done, and the results for characteristic impedance are shown in Figure 6.35. Figure 6.36 shows the track width versus gap or space dimension to result in Z = 50W.

6.24 High-Frequency Measurement of On-Chip Passives and Some Common De-Embedding Techniques

So far, we have considered inductors, transformers, and their Z-parameters. In this section, we will discuss how to obtain those Z-parameters from measurements. A typical set of test structures for measuring an inductor in a pad frame is shown in Figure 6.37. High-frequency ground-signal-ground probes will be landed on these pads so that the S-parameters of the structure can be measured. However, while measuring the inductor, the pads themselves will also be measured, and therefore two additional de-embedding structures will be required. Once the S-parameters have been measured for all three structures, a simple calculation can be performed to remove the unwanted parasitics.

The dummy open and dummy through are used to account for parallel and series parasitic effects, respectively. The first step is to measure the three structures, the device as Y_{DUT}, the dummy open as Y_{Open}, and the dummy through as $Y_{Through}$. Then, the parallel parasitic effects represented by Y_{Open} are removed as shown by

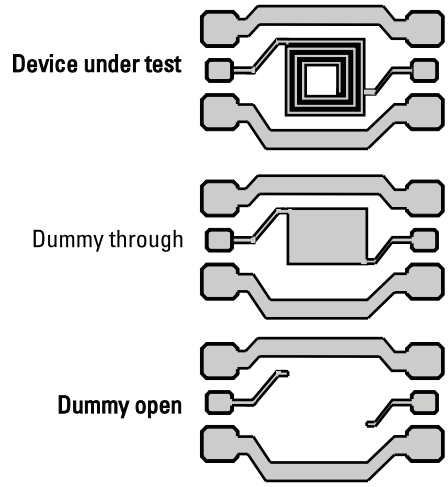

Figure 6.37 Example of a high-frequency structures used for measuring on-chip passives.

(6.39), leaving the partially corrected device admittance as Y'_{DUT} and the corrected value for the dummy through as $Y'_{Through}$.

$$Y_{DUT} = Y_{DUT} \quad Y_{Open}$$
$$Y_{Through} = Y_{Through} \quad Y_{Open} \tag{6.39}$$

The final step is to subtract the series parasitics by making use of the dummy through. Once this is done, this leaves only Z_{Device}, the Z-parameters of the device itself as shown in (6.40).

$$Z_{Device} = Z_{DUT} \quad Z_{Through} \tag{6.40}$$

where Z'_{DUT} is equal to $1/Y'_{DUT}$ and $Z'_{Through}$ is equal to $1/Y'_{Through}$.

Note that when measuring a two-port structure, Z_{11} (often displayed on the network analyzer) is not equal to the input impedance of the circuit. Also note that in general even though port two is loaded with 50W by the network analyzer,

$$Z_{in} \neq Z_{11} - 50W \tag{6.41}$$

In order to find the input impedance, one must use the formulas in Section 6.15.

6.25 Packaging

With any IC, there comes a moment of truth, a point where the IC designer is forced to admit that the design must be packaged so that it can be sold and the designers can justify their salaries. Typically, the wafer is cut up into dice with each die containing one copy of the IC. The die is then placed inside a plastic package, and the

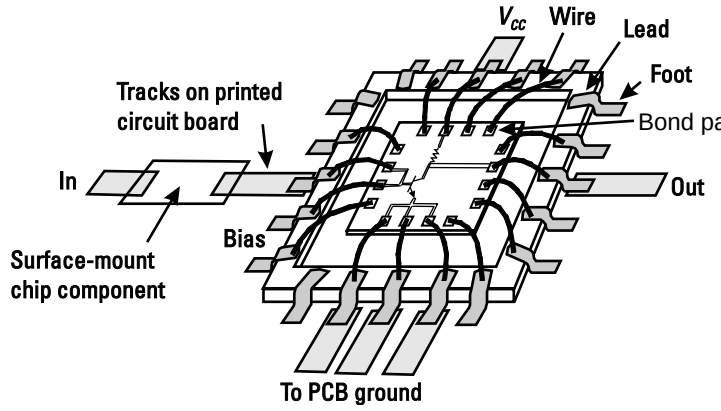

Figure 6.38 An IC in a package (delidded).

pads on the die are connected to the leads in the package with wire bonds (metal wires) as shown in Figure 6.38. The package is then sealed and can be soldered to a board.

Once the signals from the chip reach the package leads, they are entering a low-loss 50W environment and life is good. The main trouble is the impedance and coupling of the bond wires, which form inductors and transformers. For a wire of radius r a distanceh above a ground plane, an estimate for inductance is [3]

$$L \quad 0.2 \ln \frac{2h}{r} \tag{6.42}$$

where L is expressed as nanohenries per millimeter. For typical bond wires, this results in about 1 nH/mm.

For two round wires separated byd and a distanceh above a ground plane, the mutual inductance is estimated by

$$M \gg 0.1 \ln \left[1 + \left(\frac{2h}{d} \right)^2 \right] \tag{6.43}$$

where M is expressed as nanohenries per millimeter. As an example, for a pair of bond wires separated by 150µm (a typical spacing of bond pads on a chip), 1 mm from the ground plane, mutual inductance would be 0.52 nH/mm, which is a huge number. If the height is dropped to 150µm above the ground plane, then the mutual inductance of 0.16 nH/mm is still quite signi cant. Note that parallel bond wires are sometimes used deliberately in an attempt to reduce the inductance. For two inductors in parallel, each of value L_s, one expects the effective inductance to be $L_s/2$. However, with a mutual inductance of M, the effective inductance is (L_s + M)/2. Thus, with the example above, two bond wires in parallel would be expected to have 0.5 nH/mm. However, because of the mutual inductance of 0.5 nH/mm, the result is 0.75 nH/mm. Some solutions are to place the bond wires perpendicular to each other, or to place ground wires between the active bond wires (obviously of little use if we were trying to reduce the inductance of the ground connection). An-

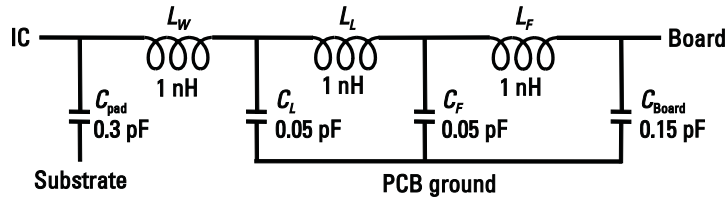

Figure 6.39 Approximate package model.

other interesting solution is to couple differential signals where the current is owing in opposite directions. This could apply to a differential circuit, or to power and ground. In such a case, the effective inductance is L(S - M)/2, which, in the above example, results in 0.25 nH/mm.

Figure 6.39 is an example of an approximate model for a 32-pin, 5-mm (3-mm die attach area) thin quad at pack (TQFP) package.

At 900 MHz, the impedances would be as shown in Figure 6.40.

The series inductor is dominant at 900 MHz. At the input and output, there is often a matching inductance, so it can simply be reduced to account for the package inductance. At the power supply, the inductance is in series with the load resistor, so gain is increased and the phase is shifted. Thus if the intended load impedance is 50W, the new load impedance is 50 - j17 or in radial terms, 52.8 Ð 18.8°. The most important effect of the package occurs at the ground pad, which is on the emitter of the common-emitter ampli er. This inductance adds emitter degeneration, which can be bene cial in that it can improve linearity, and can cause the ampli er input impedance to be less capacitive and thus easier to match. A harmful effect is that the gain is reduced. Also, with higher impedance to external ground, noise injected into this node can be injected into other circuits due to common on-chip ground connections. Note that self-resonance of this particular package limits its use to a few gigahertz.

Usually it is bene cial to keep ground and substrate impedance low, for example, by using a number of bond pads in parallel as shown in Figure 6.41. For

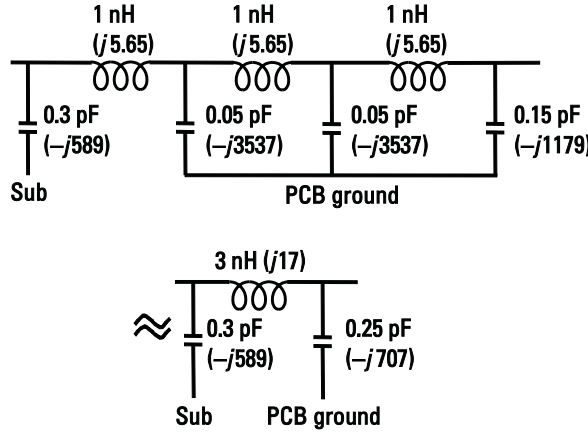

Figure 6.40 Impedances for the approximate package model of Figure 6.39 at 900 MHz.

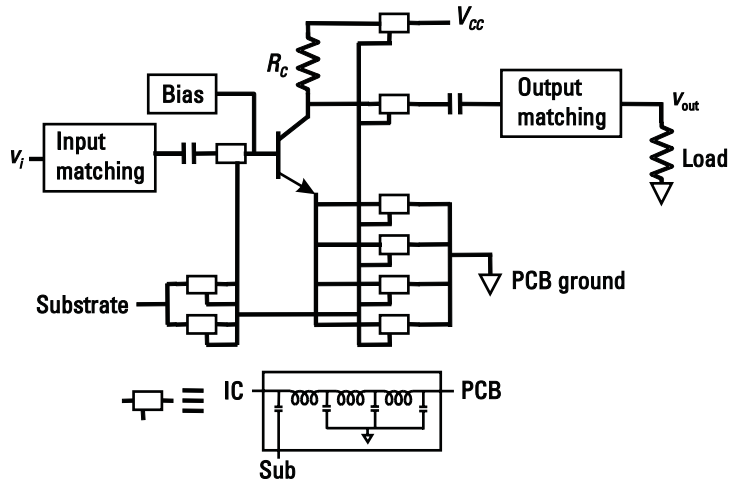

Figure 6.41 Simple ampli er with bond pad models shown.

example, with four parallel bond pads, the impedance is j4.2W. However, it must be noted that with n bond wires in parallel and close together, mutual inductance between them can increase the inductance so that inductance is not decreased by a factor of n, but by something less.

The input source and the load are referenced to the PCB ground. Multiple pads are required for the on-chip ground to have low impedance to PCB ground. Here, the emitter is at on-chip ground. The bond pads have capacitance to substrate, as do any on-chip elements as previously shown in Figure 6.6, including capacitors, inductors, transistors, and tracks. The substrate has substrate resistance, and substrate contacts are placed all over the chip and are here shown connected to several bond pads. The pads are then connected through bond wires and the package to the PCB, where they could be connected to PCB ground. While bringing the substrate connection out to the printed circuit board can be done for mixed-signal designs, for RF circuits, the substrate is usually connected to the on-chip ground.

The lead and foot of the package are over the PCB so have capacitance to the PCB ground. Note that PCBs usually have a ground plane except where there are tracks.

6.25.1 Other Packaging Techniques and Board Level Technology

Other packages are available that have lower parasitics. Examples are ip-chip and chip-on-board.

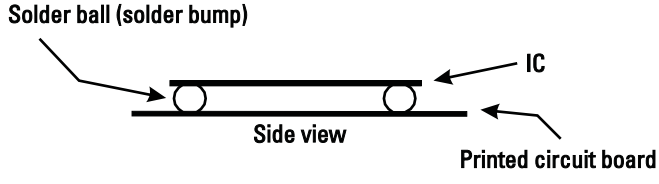

Figure 6.42 Flip-chip packaging.

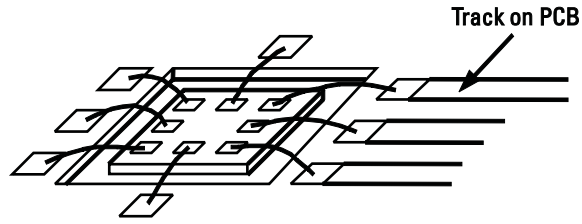

Figure 6.43 Chip-on-board packaging.

When using flip-chip packaging, a solder ball (or other conducting material) is placed on a board with a matching pattern and the circuit is connected as shown in Figure 6.42. This results in low inductance (a few tenths of a nanohenry) and very little extra capacitance. One disadvantage is that the pads must be further apart; however, the patterning of the PCB is still very fine, requiring a specialized process. Once flipped and attached, it is also not possible to probe the chip.

When using chip-on-board packaging, the chip is mounted directly on the board and bond wires run directly to the board, eliminating the package, as shown in Figure 6.43. The PCB may be recessed so the top of the chip is level with the board. This may require a special surface on the PCB (gold, for example) to allow bonding to the PCB.

Packaging has an important role in the removal of heat from the circuit, which is especially important for power amplifiers. Thermal conduction can be through contact. For example, the die may touch the metal backing. In the case of the flip-chip, thermal conduction is through solder bumps to the printed circuit board.

Recently a new type of packaging technology called low temperature cofired ceramic (LTCC) has started to be used with RF and microwave circuits. This technology allows the package level connection of multiple ICs as well as the ability to implement very fine geometry off chip passives. For example, inductors with values up to a few nanohenries with very high Q are possible. As well, custom filters can also be implemented to work with the radio, giving the designer the possibility of designing components that on-chip would be far too lossy to be practical.

References

[1] Abrie, P. L. D., RF and Microwave Amplifiers and Oscillators, Norwood, MA: Artech House 2000.

[2] Barke, E., "Line-to-Ground Capacitance Calculations for VLSI: A Comparison," IEEE Trans. on Computer-Aided-Design, Vol. 7, February 1988, pp. 195–298.

[3] Verghese, N. K., T. J. Schmerbeck, and D. J. Allstot, Simulation Techniques and Solutions for Mixed-Signal Coupling in Integrated Circuits, Norwell, MA: Kluwer, 1995.

[4] Long, J. R., "A Narrowband Radio Receiver Front-End for Portable Communications Applications," Ph.D. Dissertation, Carleton University, 1996.

[5] Mohan, S. S., et al., "Simple Accurate Expressions for Planar Spiral Inductances," IEEE J. Solid-State Circuits, Vol. 34, October 1999, pp. 1419–1424.

[6] Razavi, B., Design of Analog CMOS Integrated Circuits, New York: McGraw-Hill, 2000, Chs. 17 and 18.

[7] Danesh, M., et al., "A Q-Factor Enhancement Technique for MMIC Inductors," Proc. RFIC Symposium, 1998, pp. 183–186.

[8] Cheung, D. T. S., J. R. Long, and R. A. Hadaway, "Monolithic Transformers for Silicon RFIC Design," Proc. BCTM, September 1998,pp. 105–108.

[9] Long J. R., and M. A. Copeland, "The Modeling, Characterization, and Design of Monolithic Inductors for Silicon RF IC's," IEEE J. Solid-State Circuits, Vol. 32, March 1997, pp. 357–369.

[10] EdelsteinD. C., and J. N. Burghartz, "Spiral and Solenoidal Inductor Structures on Silicon Using Cu-Damascene Interconnects," Proc. IITC, 1998, pp. 18–20.

[11] Hisamoto, D., et al., "Suspended SOI Structure for Advanced 0.1μm CMOS RF Devices," IEEE Trans. on Electron Devices, Vol. 45, May 1998, pp. 1039–1046.

[12] Yue, C. P., and S. S. Wong, "On-Chip Spiral Inductors with Patterned Ground Shields for Si-Based RF IC's,"IEEE J. Solid-State Circuits, Vol. 33, May 1998, pp. 743–752.

[13] Craninckx, J., and M. S. J. Steyaert, "A 1.8-GHz Low-Phase-Noise CMOS VCO Using Optimized Hollow Spiral Inductors," IEEE J. Solid-State Circuits, Vol. 32, May 1997, pp. 736–744.

[14] Niknejad, A. M., and R. G. Meyer, "Analysis, Design, and Optimization of Spiral Inductors and Transformers for Si RF IC's," IEEE J. Solid-State Circuits,Vol. 33, October 1998, pp. 1470–1481.

[15] Rogers, J. W. M., J. A. Macedo, and C. Plett, "A Completely Integrated Receiver Front-End with Monolithic Image Reject Filter and VCO," IEEE RFIC Symposium, June 2000, pp. 143–146.

[16] Rogers, J. W. M., et al., "Post-Processed Cu Inductors with Application to a Completely Integrated 2-GHz VCO," IEEE Trans. on Electron Devices, Vol. 48, June 2001, pp. 1284–1287.

[17] Zolfaghari, A., A. Chan, and B. Razavi, "Stacked Inductors and Transformers in CMOS Technology," IEEE J. Solid-State Circuits,Vol. 36, April 2001, pp. 620–628.

[18] Niknejad, A. M., J. L. Tham, and R. G. Meyer, "Fully-Integrated Low Phase Noise Bipolar Differential VCOs at 2.9 and 4.4 GHz," Proc. European Solid-State Circuits Conference, 1999. pp. 198–201.

[19] van Wijnen, P. J.,On the Characterization and Optimization of High-Speed Silicon Bipolar Transistors, Beaverton, OR: Cascade Microtech, 1995.

CHAPTER 7

LNA Design

7.1 Introduction and Basic Amplifiers

The LNA is the first block in most receiver front ends. Its job is to amplify the signal while introducing a minimum amount of noise to the signal.

Gain can be provided by a single transistor. Since a transistor has three terminals, one terminal should be ac grounded, one is the input, and one is the output. There are three possibilities, as shown in Figure 7.1, each shown with a bipolar and a MOS transistor. Each one of the basic amplifiers has many common uses and each is particularly suited to some tasks and not to others. The common-emitter/source amplifier is most often used as a driver for an LNA. The common-collector/drain, with high input impedance and low output impedance, makes an excellent buffer between stages or before the output driver. The common-base/gate is often used as a cascode in combination with the common-emitter/source to form an LNA stage with gain to high frequency as will be shown. The loads shown in the diagrams can be made either with resistors for broadband operation or with tuned resonators for narrowband operation. In this chapter, the LNA with resistors will be discussed first, followed by a discussion of the narrowband LNA. As well, refinements such as feedback can be added to the amplifiers to augment their performance.

7.1.1 Common-Emitter/Source Amplifier (Driver)

To start the analysis of the common-emitter amplifier, we replace the transistor with its small-signal model, as shown in Figure 7.2(a). Z_L represents some arbitrary load that the amplifier is driving. For the common-source amplifier using a MOS transistor, the small-signal model would not have an input resistor r_π, and C_π, C_μ, υ_π, and r_b would be replaced by C_{gs}, C_{gd}, υ_{gs}, and r_g, respectively, as shown in Figure 7.2(b). Since at our frequency of interest the impedance of C_π is considerably less than the impedance of r_π, the two circuits will have similar results.

At low frequency, the voltage gain of the amplifier can be given by:

$$A_{vo} = \frac{v_o}{v_i} = \frac{r}{r_b + r} g_m Z_L \qquad \frac{Z_L}{r_e} \tag{7.1}$$

where r_e is the small signal base-emitter diode resistance as seen from the emitter. Note that $r_\pi \quad \beta r_e$ and $g_m \quad 1/r_e$. For low frequencies, the parasitic capacitances have been ignored and r_b has been assumed to be low compared to r_π.

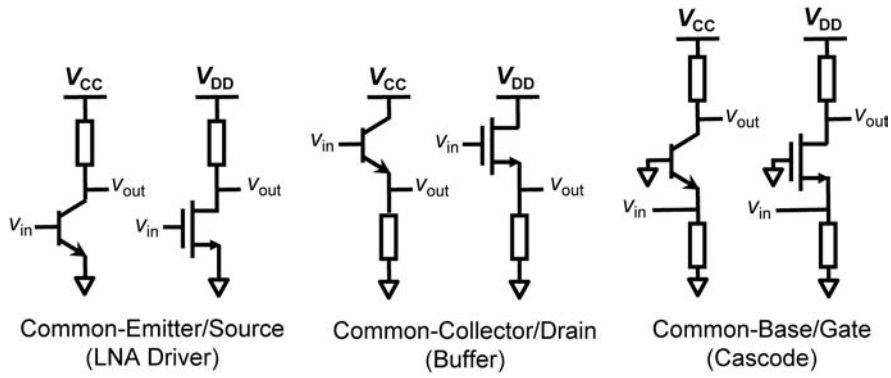

Figure 7.1 Simple transistor amplifiers.

The input impedance of the circuit at low frequencies is given by:

$$Z_{in} = r_b + r \tag{7.2}$$

In the case of a CMOS transistor, the low frequency input impedance would be given by:

$$Z_{in} = r_g + \frac{1}{sC_{gs}} \tag{7.3}$$

which will approach an open circuit at low frequency.

However, at RF, C_π will provide a low impedance across r_π, and C_μ will provide a feedback (and feedforward) path. The frequency at which the low-frequency gain is no longer valid can be estimated by using Miller's theorem to replace C_μ by two capacitors C_A and C_B, as illustrated in Figure 7.3, where C_A and C_B are:

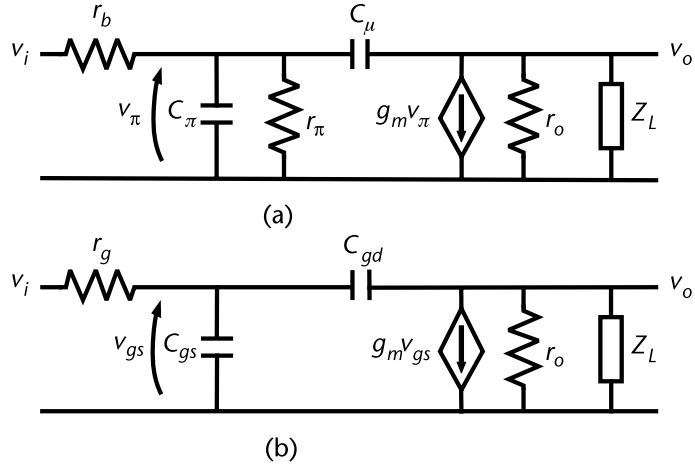

Figure 7.2 Small signal model of: (a) the common-emitter amplifier and (b) the common-source amplifier.

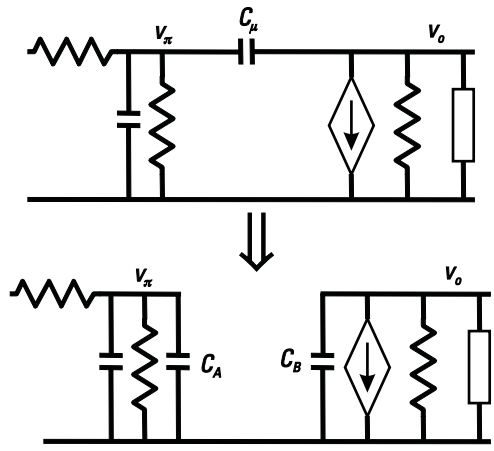

Figure 7.3 C_μ is replaced with two equivalent capacitors C_A and C_B in the common-emitter amplifier.

$$C_A = C\left(1 - \frac{v}{v}\right) = C\,(1 + g_m Z_L) \gg C\,g_m Z_L \tag{7.4}$$

$$C_B = C\left(1 - \frac{v}{v}\right) = C\left(1 + \frac{1}{g_m Z_L}\right) \gg C \tag{7.5}$$

There are now two equivalent capacitors in the circuit: one consisting of $C_A + C_\pi$ and the other consisting of C_B. This means that there are now two RC time constants or two poles in the system. The dominant pole is usually the one formed by C_A and C_π. The pole occurs at:

$$f_{P1} = \frac{1}{2\pi[r_\pi \| (r_b + R_S)][C_\pi + C_A]} \tag{7.6}$$

where R_s is the resistance of the source driving the amplifier. We note that as the load impedance decreases, the capacitance C_A is reduced and the dominant pole frequency is increased.

For the MOS transistor case, the equivalent low-frequency pole would be:

$$f_{P1} = \frac{1}{2\pi(r_g + R_S)(C_{gs} + C_A)} \tag{7.7}$$

where for the MOS case, C_A is given by:

$$C_A = C_{gd}(1 + g_m Z_L) \quad C_{gd} g_m Z_L \tag{7.8}$$

When calculating f_T, the input is driven with a current source and the output is loaded with a short circuit. This removes the Miller multiplication, and the two capacitors C_π and C_μ (or C_{gs} and C_{gd} for the MOS case) are simply connected in parallel. Under these conditions, as explained in Chapter 4, the frequency f_β where

the current gain reduces by 3 dB is given by:

$$f = \frac{1}{2\pi r_\pi (C_\pi + C_\mu)} \tag{7.9}$$

The unity current gain frequency can be found by noting that with a first-order roll-off, the ratio of f_T to f_β is equal to the low-frequency current gain β. The resulting expression for f_T is:

$$f = \frac{g_m}{2\pi(C_\pi + C_\mu)} \tag{7.10}$$

For the MOS case, the capacitors are replaced by C_{gs} and C_{gd}.

Example 7.1: Calculation of Pole Frequency

A 15 transistor, as described in Chapter 4, has the following bias conditions and properties: $I_C = 5$ mA, $r_\pi = 500$, $\beta = 100$, $C_\pi = 700$ fF, $C_\mu = 23.2$ fF, $g_m = I_C/\nu_T =$ 200mA/V, $r_b = 5$, $Z_L = 100$, and $R_s = 50$. Find the frequency f_{P1} at which the gain drops by 3 dB from its dc value.

Solution:
Since $g_m R_L = 20$,

$$C_A \quad C_\mu g_m Z_L = 23.2 \text{ fF} \times 20 = 464 \text{ fF}$$

Thus, the pole P is at a frequency of:

$$f_{P1} = \frac{1}{2\pi [500 \quad (5 + 50)][700 f + 464 f]} = 2.76 \text{ GHz}$$

Example 7.2: Calculation of Unity Gain Frequency

For the transistor in Example 7.1, compute f_β and f_T.

Solution:
Using (7.9) and (7.10), the result is:

$$f = \frac{1}{2\pi r_\pi (C_\pi + C_\mu)} = \frac{1}{2\pi \times 500\ \Omega (700 \text{ fF} + 23.2 \text{ fF})} = 440 \text{ MHz}$$

$$f_T = \frac{g_m}{2\pi (C_\pi + C_\mu)} = \frac{0.2}{2\pi (700 \text{ fF} + 23.2 \text{ fF})} = 44 \text{ GHz}$$

Knowing the pole frequency, we can estimate the gain at higher frequencies, assuming that there are no other poles present with:

$$A_v(f) = \frac{A_{vo}}{1 + j\dfrac{f}{f_{P1}}} \tag{7.11}$$

Example 7.3: Calculation of Gain of Single Pole Amplifier

For the above example, for $A_{vo} = 20$ with $f_{P1} = 2.76$ GHz, calculate the gain at 5.6 GHz.

Solution:
With $f_{P1} = 2.76$ GHz, at 5.6 GHz, the gain can be calculated to be 8.84, or 18.9 dB. This is down by about 7 dB from the low-frequency gain.

The exact expression for v_o/v_s (after about one page of algebra) is

$$\frac{v_o}{v_s} = \frac{s\,\frac{g_m}{c}}{C\,R_s\,s^2 + s\left(\frac{1}{C\,R_S} + \frac{1}{C\,Z_L} + \frac{g_m}{C}\right) + \frac{1}{C\,R_S\,CZ_L} \div} \tag{7.12}$$

where v_s is the source voltage, $R_S = R_S + r_b$, $R_{S\pi} = R_S \| r_\pi$, $C_{\mu\pi} = C_\mu$ in series with C_π, and $Z_L = Z_L \| r_o$.

As expected, this equation features a zero in the right-half plane, and real, well-separated poles, similar to that of a pole-splitting opamp [1].

Example 7.4: Calculation of Exact Poles and Zeros

Calculate poles and zeros for the transistor amplifier as in the previous example.

Solution:
Results for the previous example: 15 npn, 5 mA, Z_L =100 , $C_\mu = 23.2$ fF, $C_\pi = 700$ fF, $R_S = 50$, $r_b = 5$. Using (7.12), the results are that the poles are at 2.66 GHz and 118.3 GHz; the zero is at 1,372 GHz. Thus, the exact equation has been used to verify the original assumptions that the two poles are well separated, that the dominant pole is approximately at the frequency given by the previous equations, and that the second pole and feed-forward zero in this expression are well above the frequency of interest.

7.1.2 Simplified Expressions for Widely Separated Poles

If a system can be described by a second-order transfer function given by:

$$\frac{v_o}{v_i} = \frac{A(s\quad z)}{s^2 + sb + c} \tag{7.13}$$

then the poles of this system are given by:

$$P_{1,2} = \frac{b}{2} \pm \frac{b}{2}\sqrt{1\quad\frac{4c}{b^2}} \tag{7.14}$$

If the poles are well separated, then $4c/b^2 \ll 1$, therefore:

$$P_{1,2} \approx -\frac{b}{2} \pm \frac{b}{2}\left(1 - \frac{2c}{b^2}\right) \tag{7.15}$$

and

$$P_1 \quad \frac{c}{b} \tag{7.16}$$

$$P_2 \quad b \tag{7.17}$$

Example 7.5: Calculation of Poles and Zeros with Simplified Expressions

With the expression as above, the poles occur at 2.60 GHz and 120.94 GHz, which are reasonably close to the exact values.

7.1.3 The Common-Base/Gate Amplifier (Cascode)

The common-base (or common-gate) amplifier is often combined with the common-emitter (or common-source) amplifier to form an LNA. It can be used by itself as well. Since it has low input impedance, when it is driven from a current source, it can pass current through it with a near unity gain up to a very high frequency. Therefore, with appropriate choice of impedance levels, it can also provide voltage gain. The small-signal model for the common-base amplifier is shown in Figure 7.4 (ignoring the transistor's output impedance). The analysis for a common-gate amplifier would be identical, except for the change of names, and with no equivalent for r_π.

The current gain (ignoring C_μ and r_o) for this stage can be found to be:

$$\frac{i_{out}}{i_{in}} \quad \frac{1}{1+j\ C\ r_e} \quad \frac{1}{1+j\frac{}{}} \tag{7.18}$$

where $\omega_T = 2\pi f_T$ (see (7.10)). At frequencies below ω_T, the current gain for the stage is 1. Note that the pole in this equation is usually at a much higher frequency than the one in the common-emitter amplifier, since $r_e < r_b + R_s$. As mentioned above, the input impedance of this stage is low and is equal to $1/g_m$ at low frequencies. At the pole frequency, the capacitor will start to dominate and the impedance will drop.

This amplifier can be used in combination with the common-emitter amplifier (discussed in Section 7.1.1) to form a cascode LNA as shown in Figure 7.5. In this

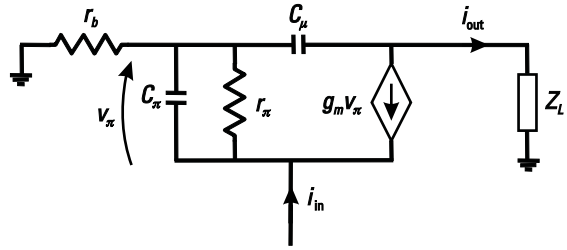

Figure 7.4 Small signal model for the common-base amplifier.

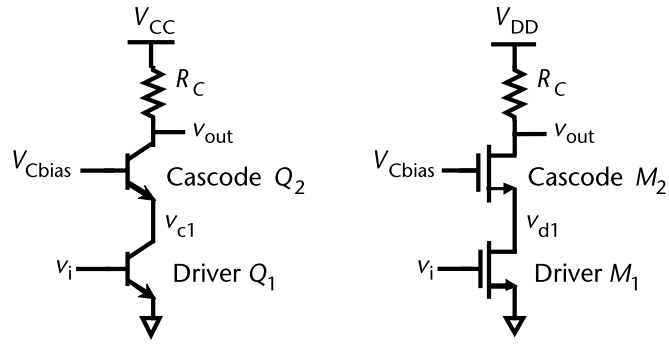

Figure 7.5 Common-base and common-gate amplifiers used as a cascode transistor in an LNA.

case, the current i_{c1} through Q_1 is about the same as the current i_{C2} through Q_2 since the common-base amplifier has a current gain of approximately 1. Then, i_{c1} $i_{c2} = g_{m1}v_i$. For the case where $R_S + r_b << r_\pi$, and v_o/v_i $-g_m R_C$, the gain is the same as for the common-emitter amplifier. However, this reduces the feedback of C_μ, and as a result, high-frequency gain increases.

In the previous section, we showed that as the load resistance on the common-emitter amplifier is reduced, the dominant pole frequency is increased. Thus, by adding the common-base amplifier as the load of the common-emitter amplifier, the impedance seen by the collector of Q_1 is r_e $1/g_{m2}$ (a low value). Thus, the frequency response of this stage has been improved by adding the common-base amplifier. For bipolar transistors, the gain of the driver stage is equal to –1 since the transconductance is directly related to the current and the current is the same in the two transistors. Thus, the Miller effect will appear to multiply C_μ by about 2. Recall that for the common-emitter amplifier, C_μ will be multiplied by about $g_m Z_L$, which would typically be much larger. For the CMOS case where the transconductance is related to both the current and the transistor size, the gain can be different from –1. If the cascode transistor is made smaller, for example, to reduce output capacitance, the gain will be larger and hence Miller multiplication will be higher.

The new estimate for the pole frequency in the cascoded common-emitter amplifier is:

$$f_{P1} = \frac{1}{2\pi[r_{\pi 1} \quad (r_{b1} + R_S)][C_\pi + 2C_\mu]} \approx \frac{1}{2\pi\left(1 + \frac{r_{\pi 1}}{R_S + r_{b1}}\right)\frac{f_T}{\beta}} \tag{7.19}$$

Example 7.6: Improving the Pole Frequency of a CE Amplifier

For the previous example with 15 npn, 5 mA, Z_L =100 , C_μ = 23.2 fF, C_π = 700 fF, R_S = 50 , r_b = 5 , estimate the pole frequency for the amplifier with a cascode transistor added.

Solution:
Without the cascode transistor, the estimated pole frequency was f_{P1} 2.7 GHz. Using (7.19), the pole frequency is 4.44 GHz.

Another advantage of the cascode amplifier is that adding another transistor improves the isolation between the two ports (very little reverse gain in the amplifier). The disadvantage is that the additional transistor adds additional poles to the system. This can become a problem for a large load resistance, leading to rapid high-frequency-gain roll-off (–12 dB/octave compared to the previous –6 dB/octave) and excess phase lag, which can cause problems if feedback is used. As well, an additional bias voltage is required, and if this cascode bias node is not properly decoupled, instability can occur. A further problem is the reduced signal swing at a given supply voltage, since the supply must now be split between two transistors instead of just one as for the simple common-emitter amplifier.

7.1.4 The Common-Collector/Drain Amplifier (Emitter/Source Follower)

The common-collector amplifier is a very useful general-purpose amplifier. It has a voltage gain that is close to 1, but has a high input impedance and a low output impedance. Thus, it makes a very good buffer stage or output stage. Also, it can be used to do a dc level shift in a circuit.

The common-collector amplifier and its small-signal model are shown in Figure 7.6. The resistor R_E may represent an actual resistor or the output resistance of a current source. Note that the Miller effect is not a problem in this amplifier since the collector is grounded. Since C_μ is typically much less than C_π, it can be left out of the analysis with little impact on the gain. The voltage gain of this amplifier is given by:

$$A_v(s) = A_{vo} \frac{1 + s/z_1}{1 + s/P_1} \div \tag{7.20}$$

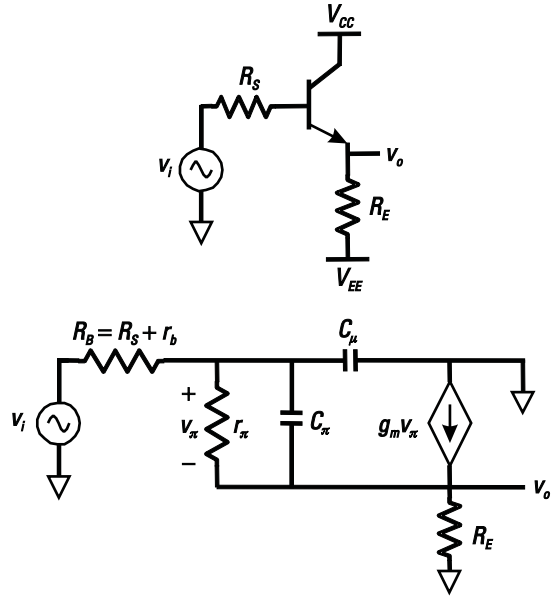

Figure 7.6 Common-collector amplifier and its small-signal model.

A_{vo} is the gain at low frequency and is given by:

$$A_{vo} = \frac{g_m R_E + R_E / r_\pi}{1 + g_m R_E + (R_B + R_E)/r_\pi} \quad \frac{g_m R_E}{1 + g_m R_E} = \frac{R_E}{1/g_m + R_E} \quad 1 \tag{7.21}$$

where $R_B = R_S + r_b$. The first approximation in (7.21) is valid only if $r_\pi >> R_B + R_E$ and this is always true for CMOS and can be valid for bipolar. The second approximation is valid only if $g_m R_E >> 1$ and while this may be valid for the bipolar implementation, it is less likely to be valid for CMOS where the transconductance is typically much lower.

The pole and zero are given by:

$$Z_1 \quad \frac{g_m}{C_\pi} = \omega_T \tag{7.22}$$

$$P_1 \quad \frac{1}{C_\pi R_A} \tag{7.23}$$

where

$$R_A = r_\pi \left\| \frac{R_B + R_E}{1 + g_m R_E} \right. \tag{7.24}$$

if $g_m R_E >> 1$, then

$$R_A \quad \frac{R_B + R_E}{g_m R_E} \tag{7.25}$$

and

$$P_1 \quad \frac{R_E}{R_E + R_B} \omega_T \tag{7.26}$$

For the CMOS circuit, the zero and pole frequencies have similar expressions with C_π replaced with C_{gs} and r_π removed.

The input impedance of this amplifier can also be determined. If $g_m R_E >> 1$, then we can use the small-signal circuit to find Z_{in}. The input impedance again ignoring C_μ is given by:

$$Z_{\text{in}} = Z_\pi + R_E(1 + g_m Z_\pi) \tag{7.27}$$

Likewise, the output impedance, not including R_E, can be found and is given by:

$$Z_{\text{out}} = \frac{r_\pi + R_B + sC_\pi r_\pi R_B}{1 + g_m r_\pi + sC_\pi r_\pi} \tag{7.28}$$

Provided that $r_\pi > R_B$ at the frequency of interest, the output impedance simplifies to:

$$Z_{\text{out}} \quad r_e \frac{1 + g_m R_B j\omega/\omega_T}{1 + j\omega/\omega_T} \tag{7.29}$$

At low frequencies, this further simplifies to:

$$Z_{\text{out}} \quad r_e \quad \frac{1}{g_m} \tag{7.30}$$

At higher frequencies if $r_e > R_B$, (recalling that $R_B = R_S + r_b$), for example, at low current levels, then $|Z_{\text{out}}|$ decreases with frequency and so the output impedance is capacitive. However, if $r_e < R_B$, then $|Z_{\text{out}}|$ increases for higher frequency and the output can appear inductive. In this case, if the circuit is driving a capacitive load, the inductive component can produce peaking (resonance) or even instability.

From (7.28) the output can be modeled as shown in Figure 7.7 where:

$$R_1 = \frac{1}{g_m} + \frac{R_s + r_b}{\beta} \tag{7.31}$$

$$R_2 = R_s + r_b \tag{7.32}$$

$$L = \frac{1}{\omega_T}(R_s + r_b) \tag{7.33}$$

For the CMOS equivalent, the output equation is somewhat simpler.

$$Z_{\text{out}} = \frac{1}{g_m}\frac{1 + sC_{gs}R_B}{1 + s\dfrac{C_{gs}}{g_m}} \tag{7.34}$$

We note that if gate resistance is small, then R_B is roughly equal to the source resistance. By setting $1/g_m$ equal to the source resistance, the pole and zero will approximately cancel out. However, in a typical application with a high driving impedance R_S, such cancellation generally cannot be done. Instead, the output impedance will also appear inductive as for the bipolar implementation.

Example 7.7: Emitter Follower Example

Calculate the output impedance for the emitter follower with a transistor, as before with 5 mA, $C_\mu = 23.2$ fF, $C_\pi = 700$ fF, $r_b = 5$. Assume that both input impedance and output impedance are 50 .

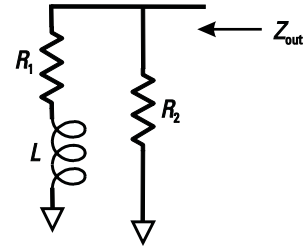

Figure 7.7 Output impedance of the common-collector amplifier.

Solution:
Solving for the various components, it can be shown that low-frequency output resistance is 5.5 and high-frequency output impedance is 55 . The equivalent inductance is 0.2 nH, the zero frequency is 4.59 GHz, and the pole frequency is 45.9 GHz.

7.2 Amplifiers with Feedback

There are numerous ways to apply feedback in an amplifier and it would be almost a book in itself to discuss them all. Only a few of the common feedback techniques will be discussed here.

7.2.1 Common-Emitter/Source with Series Feedback (Emitter/Source Degeneration)

The two most common configurations for RF LNAs are the common-emitter configuration and the cascode configuration shown in Figure 7.8. In most applications, the cascode is preferred over the common-emitter or common-source topology because it can be used up to higher frequencies (the extra transistor acts to reduce the Miller effect) and has superior reverse isolation (S_{12}). However, the cascode also suffers from reduced linearity due to the stacking of two transistors, which reduces the available output voltage swing.

Most common-emitter/source and cascode LNAs employ the use of degeneration (usually in the form of an inductor in narrowband applications) as shown in Figure 7.8. The purpose of degeneration is to provide a means to transform the real part of the impedance seen looking into the base to a higher impedance for matching purposes. This inductor also trades gain for linearity as the inductor is increased in size.

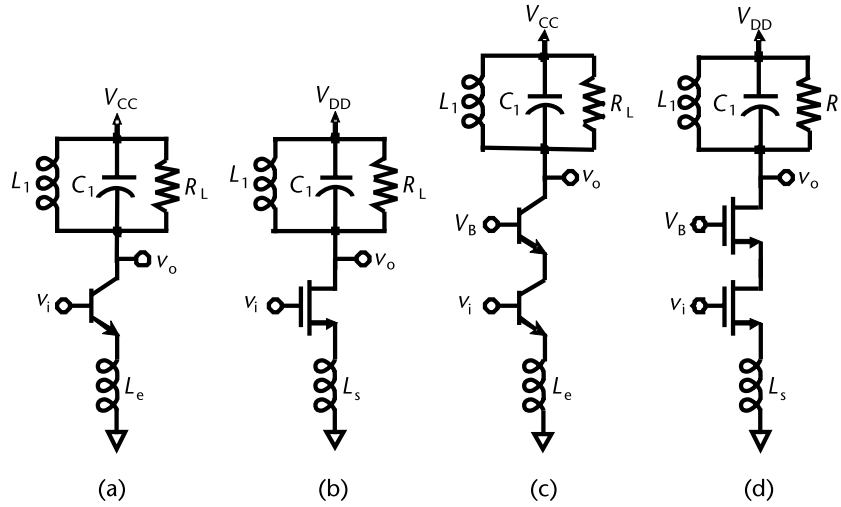

Figure 7.8 Narrowband LNAs with inductive degeneration: (a) common emitter, (b) common source, (c) cascode common emitter, and (d) cascode common source.

The gain of any of these amplifiers at the resonance frequency of the collector or drain tank, ignoring the effect of C_μ for the bipolar circuits or C_{gd} for the CMOS circuits is found with the aid of Figure 7.9 and is given by:

$$\frac{V_o}{V_i} = \frac{g_m R_L}{1 + \frac{Z_E}{Z_\pi} + g_m Z_{E\div}} \qquad \frac{R_L}{Z_E} \tag{7.35}$$

where Z_E is the impedance of the emitter or source degeneration. Here it is assumed that the impedance in the emitter or source is a complex impedance. Thus, as the degeneration becomes larger, the gain ceases to depend on the transistor parameters and becomes solely dependent on the ratio of the two impedances. This is, of course, one of the advantages of this type of feedback. Another advantage is that the circuit becomes less sensitive to temperature and process variations.

If the input impedance is matched to R_S (which would require an input series inductor), then the gain can be written out in terms of source and load resistances and f_T only. That is, the final expression is independent of the reactive component, although their values are set to ensure that the input impedance is equal to R_S. v_{out} in terms of i_x in Figure 7.9 can be given by:

$$v_{\text{out}} = \quad g_m v_\pi R_L = \quad g_m i_x Z_\pi R_L \tag{7.36}$$

Noting that i_x can also be equated to the source resistance R_s as $i_x = v_{\text{in}}/R_s$:

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{g_m Z_\pi R_L}{R_S} \tag{7.37}$$

assuming that Z_π is primarily capacitive at the frequency of interest:

$$\left|\frac{v_{\text{out}}}{v_{\text{in}}}\right| = \frac{g_m R_L}{R_S \omega_o C_\pi} = \frac{R_L \omega_T}{R_S \omega_o} \tag{7.38}$$

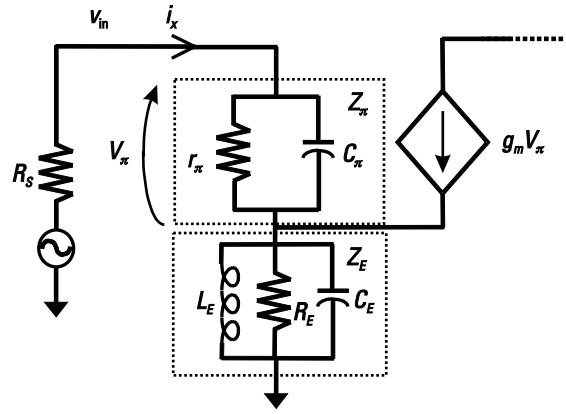

Figure 7.9 Small-signal model used to find the input impedance and gain shown for the bipolar circuits. For the CMOS circuits, r_π is absent and C_π and v_π are replaced by C_{gs} and v_{gs}, respectively.

where ω_o is the frequency of interest. Note that nodal analysis will not result in the same expression unless the input impedance is set equal to R_S and in practice this requires the addition of an input inductor. The input impedance has the same form as the common collector amplifier and is also given by:

$$Z_{\text{in}} = Z_\pi + Z_E(1 + g_m Z_\pi) \tag{7.39}$$

Of particular interest is the product of Z_E and Z_π. If the emitter impedance is inductive, then when this is reflected into the base, this will become a real resistance. Thus, placing an inductor in the emitter or source tends to raise the real part of the input impedance of the circuit, so it is very useful for matching purposes. (Conversely, placing a capacitor in the emitter or source will tend to reduce the input impedance of the circuit and can even make it negative.)

7.2.2 The Common-Emitter/Source with Shunt Feedback

Applying shunt feedback to a common-emitter or common-source amplifier is a good basic building block for broadband amplifiers. This technique allows the amplifier to be matched over a broad bandwidth while having minimal impact on the noise figure of the stage. A basic common-emitter amplifier with shunt feedback is shown in Figure 7.10. The analysis is the same for a CMOS amplifier noting that Z_π for a bipolar transistor is r_π in parallel with C_π while for CMOS the equivalent is simply C_{gs}. Resistor R_f forms the feedback and capacitor C_f is added to allow for independent biasing of the base and collector. C_f can normally be chosen so that it is large enough to be a short circuit over the frequency of interest. Note that this circuit can be modified to become a cascode amplifier if desired.

Ignoring the Miller effect and assuming C_f is a short circuit ($1/\omega C_f \ll R_f$), the gain is given by:

$$A_v = \frac{v_o}{v_i} = \frac{\dfrac{R_L}{R_f}\quad g_m R_L}{1 + \dfrac{R_L}{R_f}}\qquad \frac{g_m R_L}{1 + \dfrac{R_L}{R_f}} \tag{7.40}$$

Thus, feedback has reduced the gain compared to the original gain ($-g_m R_L$) without feedback.

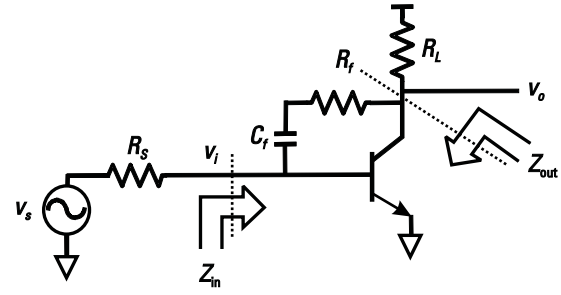

Figure 7.10 A common-emitter amplifier with shunt feedback.

The input impedance of this stage is also changed dramatically by the presence of feedback. Ignoring C_μ or C_{gd} for a CMOS amplifier, the input admittance can be computed to be:

$$Y_{\text{in}} = \frac{1}{R_f} + \frac{g_m R_L \quad \frac{R_L}{R_f}}{R_f + R_L} + \frac{1}{Z_\pi} \tag{7.41}$$

Alternatively, the input impedance can be given by:

$$Z_{\text{in}} = \frac{Z_\pi (R_f + R_L)}{R_f + R_L + Z_\pi (1 + g_m R_L)} \quad R_f \left\| Z_\pi \right\| \frac{R_f + R_L}{g_m R_L} \quad \frac{R_f + R_L}{g_m R_L} \tag{7.42}$$

This can be seen to be the parallel combination of Z_π with R_f along with a parallel component due to feedback. The last term, which is usually dominant, shows that the input impedance is equal to $R_f + R_L$ divided by the open loop gain. As a result, compared to the open-loop amplifier, the input impedance for the shunt feedback amplifier has less variation over frequency and process.

Similarly, the output impedance can be determined as:

$$Z_{\text{out}} = \frac{R_f}{1 + Z_{\text{ip}} \quad g_m \quad \frac{1}{R_f} \div} \quad \frac{R_f}{1 + g_m Z_{\text{ip}}} \tag{7.43}$$

where $Z_{\text{ip}} = R_S \quad R_f \quad Z_\pi$.

Feedback results in the reduction of the role the transistor plays in determining the gain and therefore improves linearity, but the presence of R_f may degrade the noise depending on the choice of value for this resistor.

With this type of amplifier, it is sometimes advantageous to couple it with an output buffer, as shown in Figure 7.11. The output buffer provides some inductance to the input, which tends to make for a better match. The presence of the buffer

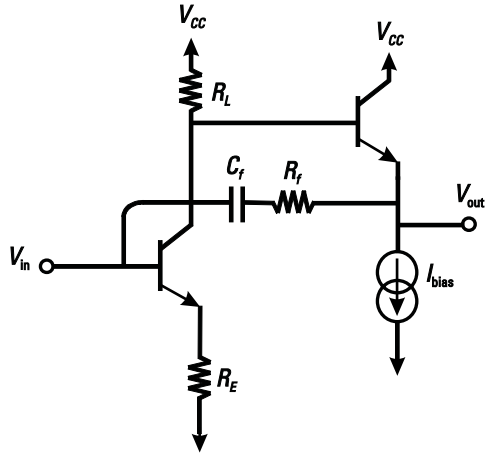

Figure 7.11 A common-emitter amplifier with shunt feedback and output CC buffer.

does change the previously developed formulas somewhat. With the addition of a buffer, the voltage gain is no longer affected by the feedback so it is approximately that of a common emitter amplifier given by $(R_L/(R_E + 1/g_m))$ minus the loss in the buffer. However, if driven from a resistor, the gain will have some dependence on the feedback resistor R_f. If the buffer is assumed to be lossless, and with a simple common-emitter amplifier at the input, that is, with $R_E = 0$, the input impedance can be determined to be:

$$Z_{\text{in}} = Z_\pi \left(1 + \frac{Z_\pi}{R_f} + \frac{g_m R_L Z_\pi}{R_f}\right)^{-1} = \frac{R_f}{1 + g_m R_L + \frac{R_f}{Z_\pi}} \approx \frac{R_f}{g_m R_L} \tag{7.44}$$

The final approximation is valid if R_f is much larger than the magnitude of Z_π. We note that it is quite common to use a nonzero R_E to improve linearity, and in such a case the input impedance will not be exactly as predicted by the equation. Simulations can be used to make any necessary adjustments.

Example 7.8: Calculation of Gain, Input, and Output Impedance with Shunt Feedback

An amplifier similar to that shown in Figure 7.10 has a load resistor of $R_L = 100$ and a source resistance of $R_S = 50$. The transistors are in a 50-GHz f_T process with 15 transistors and 5 mA for a transconductance of $g_m = 200$ mA/V. With $\beta = 100$, $r_\pi = 500$, parasitic capacitance was estimated at $C_\pi = 0.44$ pF. Assume C_μ can be ignored (the Miller effect can be minimized by using a cascode structure). At 900 MHz, compare gain, input, and output impedance with feedback resistor at $R_f \to$ and $R_f = 500$.

Solution:

Without feedback from the output to the base, the gain is $g_m R_L$ is 20 or 26 dB. From input source to output including the effect of the source resistance and Z_π, calculated as 196–*j*244 (or 313∠ –51.2°), the gain A_v is equal to:

$$A_v = (\ g_m R_L)\frac{Z_\pi}{R_s + Z_\pi} = \ 20\frac{196 \quad j244}{50 + 196 \quad j244} = 17.95 \quad j2.03$$

$$= 18.07 \qquad 6.45° \text{ (or 23.3 dB)}$$

Input impedance is given by Z_π, calculated as 196 *j*244 (or 313∠ – 51.2°). The output impedance for this simple example is infinity. However, for a real circuit, the output impedance would be determined by the transistor, integrated circuit, and package parasitic impedance.

Now for the case where the feedback is 500 , from (7.40), the voltage gain is:

$$A_v = \frac{v_o}{v_i} \quad \frac{0.2 \quad 100}{1 + \frac{100}{500}} = \frac{20}{1.2} = 16.67 \quad \text{(or 24.4 dB)}$$

(Note the exact expression from the same equation would have resulted in 16.5, so the approximate expression is sufficient.) Thus, the gain has been reduced from 26 dB to 24.4 dB.

From the source, the gain is reduced much more since the input impedance is much reduced so there is a lot of attenuation. The gain from the source, instead of from v_i, can be shown to be just over 15 dB.

Thus, gain is reduced from 23.8 dB to about 15 dB by feedback.

$$Z_{\text{in}} \quad R_f \left\| Z_\pi \right\| \frac{R_f + R_L}{g_m R_L} = 500 \left\| (196 \quad j244) \right\| \frac{500+100}{20} \div = 28.16 \quad j1.98$$

$$Z_{\text{out}} \quad \frac{R_f}{1+g_m Z_{\text{ip}}} = \frac{500}{1+0.2 \quad (41.22 \quad j4.27)} = 53.63 + j4.96$$

where $Z_{\text{ip}} = R_s \ R_f \ Z_s = 50 \ 500 \ (196 \quad j244) = 41.22 \quad j4.27$.

We note that at low frequency without feedback, Z_{in} would have been 500 due to r_π, while with feedback, it can be shown to be 27 (assuming the feedback capacitor is large enough so that feedback is still being applied), so the impedance is much more steady across frequency.

In Section 7.9, we will conclude the chapter with a major example of a broadband amplifier design, including considerations of noise and linearity. Noise and linearity are topics that will be discussed in some of the following sections.

7.3 Noise in Amplifiers

When the signal is first received by the radio, it can be quite weak and can be in the presence of a great deal of interference. The LNA is the first part of the radio to process the signal and it is therefore essential that it amplify the signal while adding a minimal amount of additional noise to it. Thus, one of the most important considerations when designing an LNA is the amount of noise present in the circuit. The following sections discuss this important topic.

7.3.1 Input Referred Noise Model of the Bipolar Transistor

Note that the following two sections contain many equations, which may be tedious to some readers. Reader discretion is advised. Those who find equations disturbing can choose to go directly to Section 7.3.4.

In Chapter 2, we made use of an idealized model for an amplifier with two noise sources at the input. If the model is to be applied to an actual LNA, then all the noise sources, as discussed in Chapter 4, must be written in terms of these two input-referred noise sources, as shown in Figure 7.12. Starting with the model shown in Figure 7.12(a) and assuming that the emitter is grounded with base input and collector output, the model may be determined with some analysis.

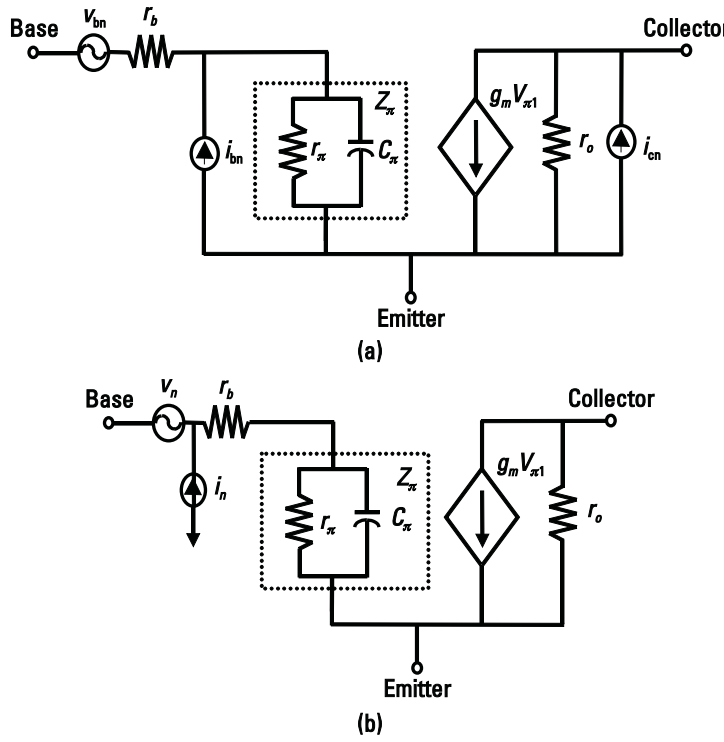

Figure 7.12 Noise model for (a) the bipolar transistor and (b) the equivalent input referred noise model.

When the input is shorted in Figure 7.12(b), v_n is the only source of noise in the model, and assuming that r_b is small enough to have no effect on the gain, the output noise current $i_{\text{on\_tot}}$ would be:

$$\overline{i_{\text{on\_tot}}^2} = \overline{v_n^2} g_m^2 \tag{7.45}$$

If instead the actual noise sources in the model are used as in Figure 7.12(a), then the output noise current can also be found. In this case, the effect of the base shot noise is assumed to be shorted out (i.e., r_b is small compared to the input impedance of the transistor). That leaves the collector shot noise and base resistance noise sources in the model to be accounted for.

$$\overline{i_{\text{on\_tot}}^2} = \overline{v_{bn}^2} g_m^2 + \overline{i_{cn}^2} \tag{7.46}$$

We wish to make these two models equivalent, so we set (7.45) equal to (7.46) and solve for v_n, giving:

$$\overline{v_n^2} = \frac{2qI_c}{g_m^2} + 4kTr_b \tag{7.47}$$

where $\overline{v_{bn}^2} = 4kTr_b$ and $\overline{i_{cn}^2} = 2qI_c$. Now if the input is open circuited in Figure 7.12(b), then only i_n can have any effect on the circuit. In this case the output noise is:

$$\overline{i_{\text{on\_tot}}^2} = \overline{i_n^2} Z_\pi^2 g_m^2 \tag{7.48}$$

Similarly for the model in Figure 7.12(a):

$$\overline{i_{\text{on\_tot}}^2} = \overline{i_{bn}^2} Z_\pi^2 g_m^2 + \overline{i_{cn}^2} \tag{7.49}$$

Now solving (7.48) and (7.49) for i_n gives:

$$\overline{i_n^2} = 2qI_B + \frac{2qI_C}{g_m^2} Y_\pi^2 \tag{7.50}$$

where $\overline{i_{bn}^2} = 2qI_B$ and $\overline{i_{cn}^2} = 2qI_c$.

7.3.2 Noise Figure of the Common-Emitter Amplifier

Now that the equivalent input referred noise model has been derived, it can be applied to the results in Chapter 2 so that the optimum impedance for noise can be found in terms of transistor parameters.

The input-referred noise current has two terms. One is due to base shot noise and one is due to collector shot noise. Since collector shot noise is present for both v_n and i_n, this part of the input noise current is correlated with the input noise voltage, but the other part is not. Thus:

$$\overline{i_c^2} = \frac{2qI_C}{g_m^2} Y_\pi^2 \tag{7.51}$$

$$\overline{i_u^2} = 2qI_B \tag{7.52}$$

Likewise, in the case of the v_n:

$$\overline{v_c^2} = \frac{2qI_C}{g_m^2} \tag{7.53}$$

$$\overline{v_u^2} = 4kTr_b \tag{7.54}$$

The correlation admittance Y_c can be determined (see Section 2.2.5). It will be assumed that at the frequencies of interest the transistor looks primarily capacitive.

$$Y_c = \frac{i_c}{v_c} = \sqrt{\frac{\frac{2qI_C}{g_m^2} Y_\pi^2}{\frac{2qI_C}{g_m^2}}} = Y_\pi = j\omega C_\pi \tag{7.55}$$

where it is assumed that r_π is not significant. Explicitly:

$$G_c \quad 0 \tag{7.56}$$

$$B_c = \omega C_\pi \tag{7.57}$$

Thus, the correlation admittance is just equal to the input impedance of the transistor.

R_c, R_u, and G_u can also be written down directly.

$$R_c = \frac{\overline{v_c^2}}{4kT} = \frac{2qI_C}{4kTg_m^2} = \frac{v_T}{2I_C} = \frac{1}{2g_m} \tag{7.58}$$

$$R_u = \frac{\overline{v_u^2}}{4kT} = \frac{4kTr_b}{4kT} = r_b \tag{7.59}$$

$$G_u = \frac{\overline{i_u^2}}{4kT} = \frac{2qI_B}{4kT} = \frac{I_C}{2V_T\beta} = \frac{g_m}{2\beta} \tag{7.60}$$

Using these equations, an explicit expression for the noise factor can be written in terms of circuit parameters by making use of the general expression for F in terms of the above parameters as shown in Chapter 2 in (2.28). The following expressions were used to simplify the results

$$\omega_T \quad \frac{g_m}{C_\pi}, \quad g_m = \frac{I_C}{V_T}, \quad r_e \quad \frac{1}{g_m} \tag{7.61}$$

The resulting expression for F is

$$F = 1 + \frac{R_s}{2r_e\beta} + \frac{r_e}{2R_s} + \frac{R_s}{2r_e}\left(\frac{f}{f_T}\right)^2 + \frac{r_b}{R_s} \tag{7.62}$$

Here it is assumed that the source resistance has no reactive component.

These equations also lead to expressions for G_{opt} and B_{opt} making use of (2.29) and (2.30) in Chapter 2:

$$G_{\text{opt}} = \sqrt{\frac{g_m^2}{\beta(1+2g_mr_b)} + \frac{2g_mr_b(\omega C_\pi)^2}{(1+2g_mr_b)^2}} \tag{7.63}$$

$$B_{\text{opt}} = \frac{\omega C_\pi}{1+2g_mr_b} \tag{7.64}$$

We note that G_{opt} varies directly with device size (since I_C, g_m, and C_π are proportional to transistor size and r_b is inversely proportional to transistor size). Thus, device size can be adjusted to set G_{opt} equal to 1/50 . Thus for that size of device, driving it with 50 would give the best noise figure. In the next section

we will see how to set the actual input impedance so that it is also equal to 50 , making it feasible to match both power and noise simultaneously.

The expression for B_{opt} can be simplified if r_b is small compared to $1/2g_m$. This would be a reasonable approximation over most normal operating points:

$$B_{\text{opt}} = \omega C_\pi \tag{7.65}$$

Thus, it can be seen that the condition for maximum power transfer (resonating out the reactive part of the input impedance) is the same as the condition for providing optimal noise matching.

Hence, the method in the next section can be used for matching an LNA.

7.3.3 Noise Model of the CMOS Transistor

A similar procedure can be followed as for the bipolar transistor with similar results. For bipolar transistors, the noise sources were collector shot noise current, base shot noise current, and base resistance noise voltage. For the CMOS transistor, similar noise sources are the drain channel noise current, gate induced noise current, and gate resistance noise voltage as shown in Chapter 4 and repeated in Figure 7.13.

By going through the same steps, or by directly comparing noise at the input due to the various noise sources, the following expression can be derived for noise figure of a simple NMOS common source amplifier:

$$F = 1 + \frac{r_g + r_{LG}}{R_s} + \frac{\gamma}{g_m R_S} + \frac{\delta g_m R_s}{5} \left(\frac{\omega}{\omega_T} \div \right)^2 \tag{7.66}$$

where r_{LG} is the series resistance of inductor L_G.

This derivation has been simplified by assuming that the common source configuration has a high input impedance; hence, drain channel noise is reflected back to the input using voltage only. After impedance matching as discussed in the following sections, the above equations for noise are no longer quite correct as the gain is modified by the impedance matching circuit. A modified equation for noise

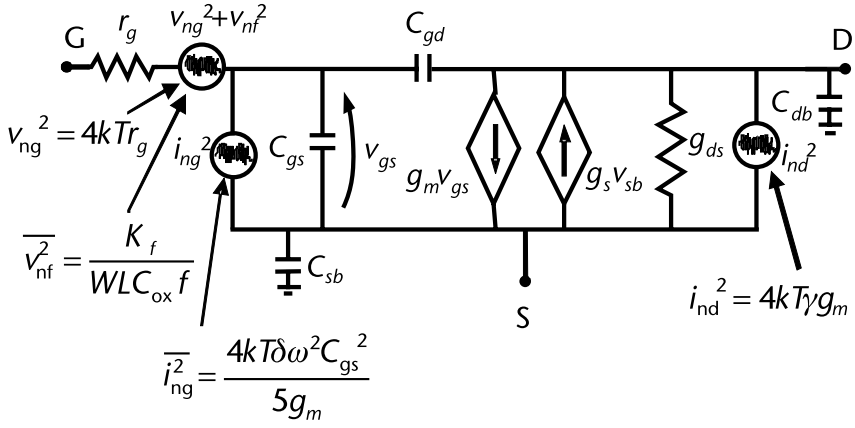

Figure 7.13 Noise model for the CMOS transistor.

will be shown in the CMOS design example; however, it can be noted that the results are not that different. It can also be noted that in practice at low gigahertz frequencies, the gate induced noise term, the last term in (7.66), can often be left out to simplify the calculations.

7.3.4 Input Matching of LNAs for Low Noise

Since the LNA is the first component in the receiver chain, the input must be matched to be driven by 50 . Many methods for matching the input using passive circuit elements are possible with varying bandwidths and degrees of complexity, many of which have already been discussed; however, one of the most elegant is described in [2]. This method requires two inductors to provide the impedance and noise match for the LNA as shown in Figure 7.14.

Starting with (7.39), and using the component names for the bipolar transistor, the input impedance for this transistor (assuming that the Miller effect is not important and that r_π is not significant at the frequency of interest) is:

$$Z_{\text{in}} = \frac{j}{\omega C_\pi} + j\omega L_e + \frac{g_m L_e}{C_\pi} + j\omega L_b \tag{7.67}$$

Note that to be matched, the real part of the input impedance must be equal to the source resistance R_s so that:

$$\frac{g_m L_e}{C_\pi} = R_s \tag{7.68}$$

Therefore:

$$L_e = \frac{R_s C_\pi}{g_m} = \frac{R_s}{\omega_T} \tag{7.69}$$

Note that C_μ has been ignored. If it were considered, the value of the capacitance would be larger than C_π, and therefore a larger inductor would be required to perform the match.

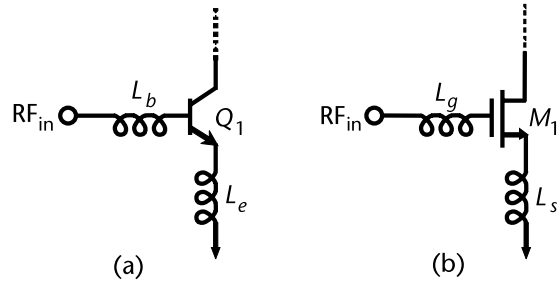

Figure 7.14 LNA driver transistor with two inductors to provide power and noise matching: (a) with bipolar transistor and (b) with MOS transistor.

Also, the imaginary part of the input impedance must equal zero. Therefore:

$$L_b = \frac{1}{C_\pi \omega^2} \quad \frac{R_s C_\pi}{g_m} \tag{7.70}$$

Making use of the above analysis, as well as the discussions on noise so far, the following method for simultaneously matching an LNA for power and noise was created. It is outlined in the following steps.

1. Find the current density in the process that will provide the lowest minimum NF and set the current density in the transistor to be this value, regardless of the size of the device. The minimum NF for a process can be found from device measurements, but for the circuit designer it can be determined by the use of simulators such as ADS, SPICE, or Spectre.
2. Once the current density is known, then the length of the transistor l_e (equivalently, the transistor emitter area, since the width is constant) should be chosen so that the real part of the optimum source impedance for lowest noise figure is equal to 50 . The current must be adjusted in this step to keep the current density at its optimal level determined in step 1.
3. Size L_e, the emitter degeneration inductor, so that the real part of the input impedance is 50 . The use of inductive degeneration will tend to increase the real part of the input impedance. Thus, as this inductor is increased, the impedance will (at some point) have a real part equal to 50 .
4. The last step in the matching is simply to place an inductor in series with the base L_b. Without this inductor, the input impedance is capacitive due to C_π. This inductor is sized so that it resonates with L_e and C_π at the center frequency of the design. This makes the resultant input impedance equal to 50 with no additional reactive component.

This technique has several advantages over other matching techniques. It is simple and requires only one additional matching component in series with the input of the transistor. It produces a relatively broadband match, and it achieves simultaneous noise and power matching of the transistor. This makes it a preferred method of matching.

There are many instances where this method cannot be applied without modification. For some designs, power constraints may not accommodate the necessary current to achieve the best noise performance. In this case, the design could proceed as before except with a less than optimum current density. In other cases, the design may not provide the necessary gain required for some applications. In this case, more current may be needed, or some other matching technique that does not require degeneration may have to be employed. As well, linearity constraints may demand a larger amount of degeneration than this method would produce.

Example 7.9: Simultaneous Noise and Power Matching with a Bipolar Transistor

Design an LNA to work at 5 GHz using a 1.8-V supply with the simultaneous noise and power matching technique discussed in this chapter. For the purpose of the

example design, a simple buffer so that the circuit can drive 50 . Assume that a 50-GHz, 0.5-μm SiGe technology is available.

Solution:
At 1.8V it is still possible to design an LNA using a cascode configuration. The cascode transistor can have its base tied to the supply. A simple output buffer that will drive 50 quite nicely is an emitter follower. Thus, the circuit would be configured as shown in Figure 7.15. In this case V_{bias} will be set to 1.8V. The buffer should be designed to accommodate the linearity required by the circuit and drive 50 , but for the purposes of this example we will set the current I_{bias} at 3 mA. This will mean that the output impedance of the circuit will be roughly 8.3 and there will be very little voltage loss through the follower. Note that in industry common-emitter buffers are typically not used to drive off-chip loads at this frequency for reasons of stability and short circuit protection. However, if the LNA drives an on-chip mixer, this should be fine.

Next the size of the tank capacitor and inductor must be set. We will choose an inductance of 1 nH. If the Q of this inductor is 10, then the parallel resistance of the inductor at 5 GHz is 314 . The inductor can be adjusted later to either move the gain up or down as required. With 1 nH of inductance, this means that a capacitance of 1 pF is required to resonate at 5 GHz.

Now the current density that results in lowest NF_{min} must be determined. Before we go to the simulator, we can make some estimates using the transistors described in Chapter 4; for example, the transistor labeled as 15 had a peak f_T of about 60 GHz at 8 mA, so it is close to the transistors being used in this example. It has been observed that for a bipolar transistor, the minimum noise figure typically occurs at about one-eighth of the current density required for optimal f_T; thus, a starting current of 1 mA is used. From Figure 4.9 at 1 mA, these transistors have

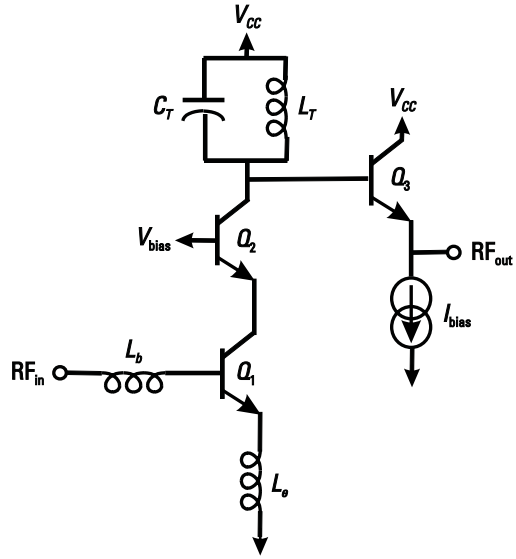

Figure 7.15 Cascode LNA with output buffer.

an f_T of about 34 GHz. Using the equation for f_T, and noting that at 1 mA, g_m is 0.04A/V, an estimate for C_π is

$$C_\pi \quad \frac{g_m}{\omega_T} = \frac{0.04}{2\pi \quad 34G} = 187\text{fF}$$

Also noting that the 15 transistor in Chapter 4 had an r_b of 5 , $g_m r_b$ is equal to 0.2, G_{opt} can be estimated from (7.63)

$$G_{\text{opt}} = \sqrt{\frac{{g_m}^2}{\beta(1+2g_m r_b)} + \frac{2g_m r_b(\omega C_\pi)^2}{(1+2g_m r_b)^2}} \div = \sqrt{\frac{0.04^2}{100(1.4)} + \frac{2 \quad 0.02(2\pi \quad 5G \quad 187f)^2}{(1.4)^2}}$$

$$= \sqrt{11.43\mu + 7.04\mu} = \sqrt{18.47\mu} = 0.00430$$

This translates into $R_{\text{opt}} = 1/G_{\text{opt}}$ 233 , but we need $G_{\text{opt}} = 1/50 = 0.02\text{A/V}$. We need G_{opt} to be about 4.65 times bigger, and since G_{opt} is proportional to size and current, we need to scale the transistor and the current by 4.65 times. The resulting transistor is 69.75x, the current I_C is 4.65 mA, C_π is 870 fF, g_m is 0.186 A/V, and r_b is 1.08 .

This paper exercise could be continued to determine matching components and to predict the noise figure, but instead we will go onto simulations and continue from there. To determine the current density for optimal noise, the current in Q_1 and Q_2 is swept. Since we do not have a handy conversion between the transistor sizes in Chapter 4 and dimensions in microns, we start rather arbitrarily with a 20-μm emitter length. The results are shown in Figure 7.16 and show that for the technology used here, a current of 3.4 mA is optimal, or about 170 μA/μm.

Next we sweep the transistor length, keeping the transistor current density constant to see where the real part of the optimal noise impedance is 50 . The results of this simulation are shown in Figure 7.17. From this graph, it can be seen that an emitter length of 27.4 μm will be the length required. This corresponds to a current

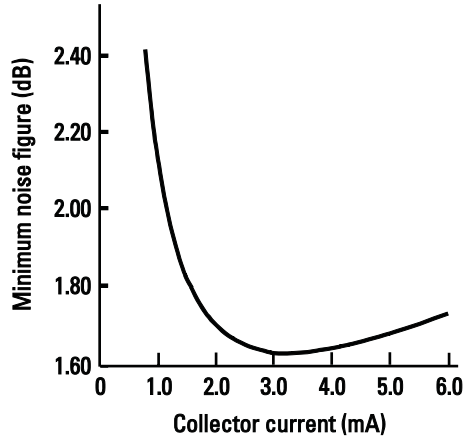

Figure 7.16 Minimum noise figure plotted versus bias current for a 20-μm transistor.

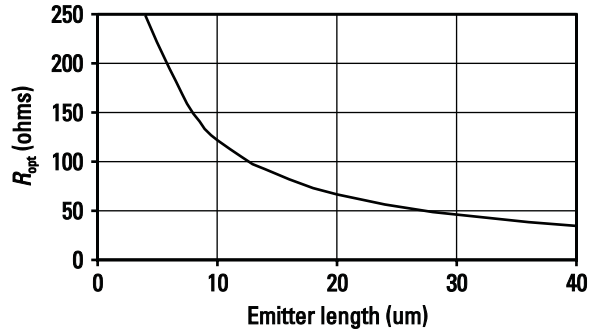

Figure 7.17 R_{opt} plotted versus emitter length with current density set for low noise.

of 4.6 mA, very nearly the same as the predicted results. Now the transistor size and current are determined.

From a dc simulation the C_π of the transistor may be found and in this case was 742.5 fF. Since the current is 4.6 mA, the g_m is 184 mA/V and thus the f_T is 39.4 GHz. This is probably a bit optimistic, but this can be used to compute initial guesses for L_e and L_b.

$$L_e = \frac{R_s}{\omega_T} = \frac{50}{248\ \text{Grad/s}} = 200\ \text{pH}$$

$$L_b = \frac{1}{C_\pi \omega^2} \quad \frac{R_s C_\pi}{g_m} = \frac{1}{(742.5\ \text{fF})(2\times\pi\times 5\ \text{GHz})^2} \quad \frac{50 \quad (742.5\ \text{fF})}{184\ \text{mA/V}} = 1.16\ \text{nH}$$

These two values were refined with the help of the simulator. In the simulator $L_e = 290$ pH and $L_b = 1$ nH was found to be the best choice.

Once these values were chosen, a final set of simulations on the circuit was performed.

First, the S_{11} of the circuit was plotted to determine that the matching was successful. A plot of S_{11} is shown in Figure 7.18. Note that the circuit is almost

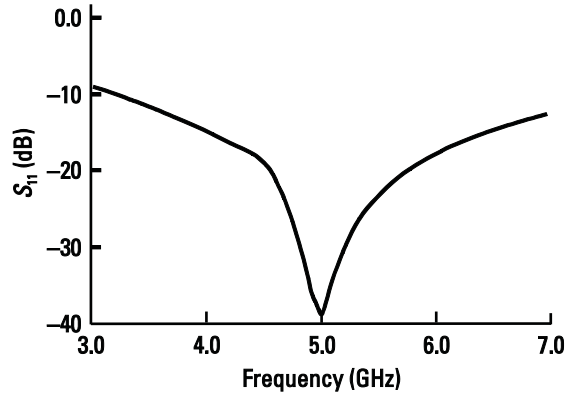

Figure 7.18 Plot showing the input matching for the LNA.

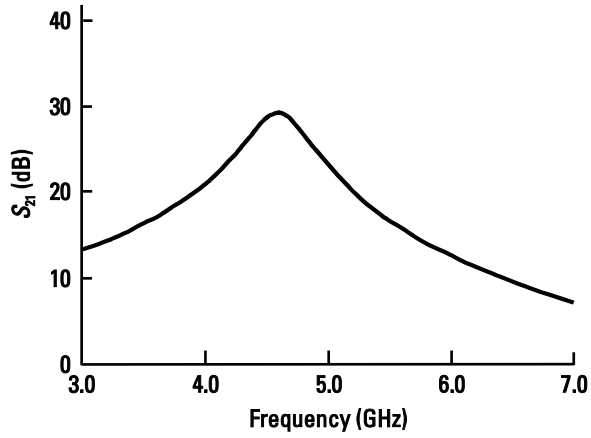

Figure 7.19 Plot showing the gain of the LNA.

perfectly matched at 5 GHz as designed. Of course in practice, loss from inductors as well as packaging and bond wires would never allow such perfect results in the lab.

The gain is plotted in Figure 7.19. Note that it peaks about 400 MHz lower than initially calculated. This is due to the capacitance of the transistor Q_2 and the output buffer transistor Q_3. This could be adjusted by reducing the capacitor C_T until the gain is once more centered properly. The gain should have a peak value given by (7.38), which in this case would have a value of:

$$\left|\frac{v_{\text{out}}}{v_{\text{in}}}\right| = \frac{R_L \omega_T}{R_S \omega_o} = \frac{314 \quad (2\times\pi\times39\times4 \text{ GHz})}{50 \quad (2\times\pi\times5 \text{ GHz})} = 33.9 \text{ dB}$$

minus the loss in the buffer. The gain was simulated to have a peak value of about 29 dB, which is close to this value. If this gain is found to be too high, it can be reduced by reducing L_T or adding resistance to the resonator.

Finally, the noise figure of the circuit was plotted as shown in Figure 7.20. The design had a noise figure of 1.76 dB at 5 GHz, which is very close to the minimum

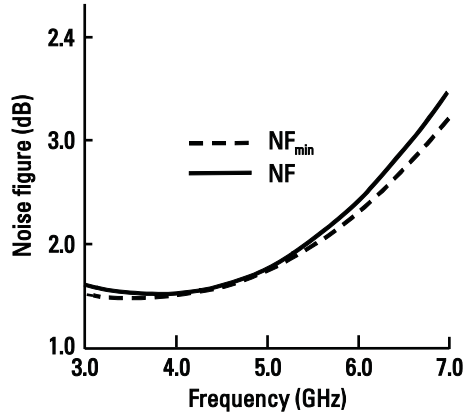

Figure 7.20 Plot showing the noise figure compared to the minimum noise figure for the design.

achievable noise figure of 1.74 dB ($F = 1.49$) showing that we have, in fact, a noise match for the circuit at 5 GHz.

A calculation can be done for noise to compare to simulated values using (7.62)

$$F = 1 + \frac{g_m R_s}{2\beta} + \frac{1}{2g_m R_s} + \frac{g_m R_s}{2}\left(\frac{f}{f_T}\right)^2 + \frac{r_b}{R_s}$$

$$= 1 + 0.058 + 0.054 + 0.074 + 0.022 = 1.208 \qquad 0.82 \text{ dB}$$

where we have used $\beta = 80$, $g_m = 0.184$, $R_s = 50$, $f = 5$G, and $f_T = 39.4$ GHz.

Thus, simulated noise is significantly higher than predicted noise, although in terms of F the discrepancy is only 0.22, which is about 15%. When the overall noise is very low, as in this example, more noise sources become important. There are at least three other sources of noise not accounted for by this formula: (1) noise in the cascode transistor, (2) noise in the buffer transistor, and (3) noise generated by the equivalent load resistor (although with 30 dB of gain, this will not be very significant). The exact reasons for any discrepancy can be determined through a noise summary.

We note that the same technique will work for CMOS amplifiers with the appropriate change of names for component values. If the absolute minimum noise is sought, the power dissipation will tend to be high, but since noise remains close to its minimum value over a range of bias currents, considerably lower power dissipation can be achieved with only a slight penalty of noise. Also, note that for CMOS, the transistor g_m is not as easily calculated and instead is done through characterization of a standard test transistor.

Example 7.10: Simultaneous Noise and Power Matching with MOS Transistor

Design an LNA to work at 5.2 GHz using a 1.2-V supply with the simultaneous noise and power matching technique discussed in this chapter. As in the bipolar example, design a simple buffer so that the circuit can drive 50 . Assume that a 0.13-μm CMOS technology is available.

Solution:

As with the bipolar example, a cascode configuration will be used to minimize the Miller effect and to allow easier matching. Because of the low threshold voltage, a cascade design is straightforward with a 1.2-V power supply. The cascode transistor gate is tied to the power supply. A source follower is used as the buffer. As with the bipolar case, the source follower is susceptible to instability if driving capacitive loads (it is transformed into negative resistance at its gate); however, if the load is intended to be on chip and its impedance is well known, this should be fine. The complete circuit is shown in Figure 7.21. For this example, the buffer size and current were set simply to provide 50 output impedance, but in practice, if linearity is the limiting factor, buffer current would need to be increased and the resulting output impedance would be lower. If, on the other hand, the LNA is driving further

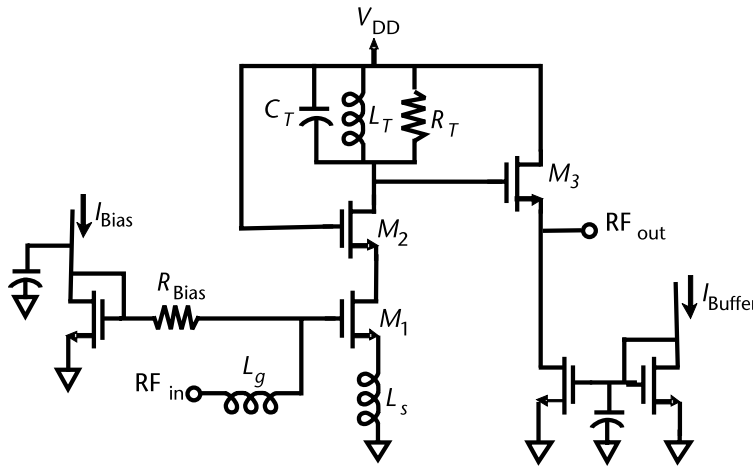

Figure 7.21 Cascode CMOS LNA with output buffer and with current sources shown.

in-chip components, it may be possible to operate with much higher output impedance and hence much lower current.

Initially, we start with the same transistor characterized in Chapter 4, with a 20-micron gate width. Its minimum noise figure NF_{min} versus transistor bias current is shown in Figure 7.22 with the minimum value for NF_{min} occurring at about 2 mA. The real part of the optimum source impedance to achieve this noise figure was determined to be about 500 . Thus, to achieve a minimum noise figure at 50 , it is estimated that the transistor size needs to be increased 10 times and the optimum will occur at a current of about 20 mA. Simulating a 200-μm transistor as in Figure 7.22(a) and extracting parameters show that the optimal source resistance is just under 50 (see Figure 7.22(b)), minimum NF_{min} is about 0.6 dB (see also Figure 7.23(a)), and the current required for the minimum noise figure is about 12 mA (see Figure 7.23(b)).

Figure 7.22(b) and Figure 7.23 were generated by sweeping transistor sizes and determining the optimum source resistance, the minimum NF_{min}, and the bias current to achieve this minimum NF_{min}. The bias current of about 12 mA to achieve a minimum NF_{min} with an optimal source resistance of 50 is quite high, but as

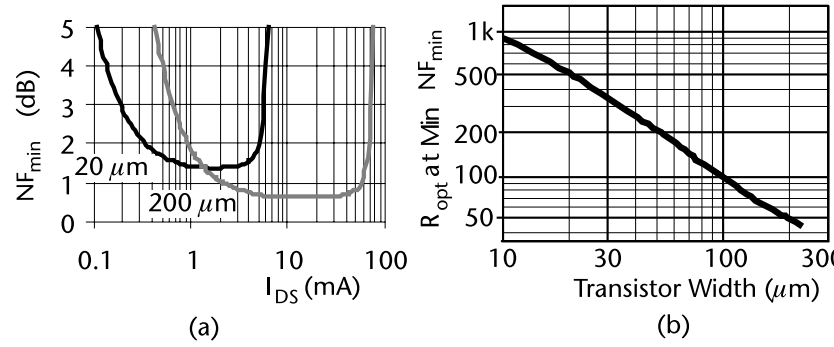

Figure 7.22 (a) Minimum noise figure for 20-micron and 200-micron transistors. (b) Real part of optimum source impedance for minimum noise (R_{opt}).

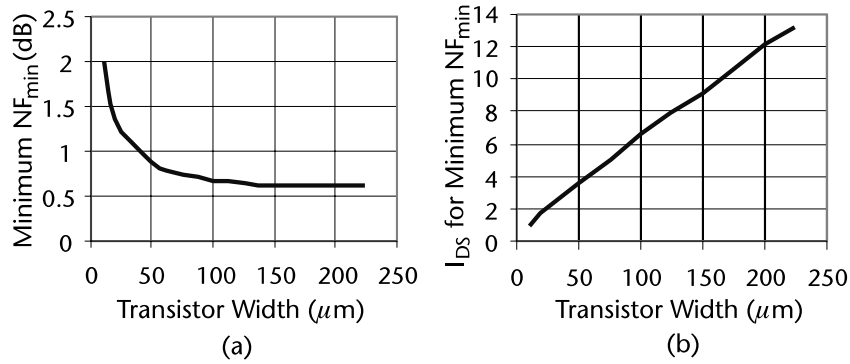

Figure 7.23 (a) Minimum NF_{min} versus transistor size. (b) Current required to achieve minimum NF_{min}.

seen from Figure 7.22(a), for a 200-μm transistor, there is very little noise penalty for decreasing the current down to about 2 mA, so this value was selected. At 2 mA, the transistor has a g_m of about 40 mA/V, an input C_{gs} of about 230 fF, and an f_T of about 28 GHz. We note that at 2 mA rather than at the optimal 12 mA, the optimal source impedance is no longer exactly 50 , so further refinements would be possible.

To size the inductors, the following equations were used. L_S was sized from the MOS equivalent of (7.69):

$$L_s = \frac{C_{gs}R_s}{g_m} \quad \frac{R_s}{\omega_T} = \frac{50}{2\pi \times 28G} = 284 \text{ pH}$$

It should be noted that the f_T for a single transistor was used to design a cascade amplifier, but the results are still fairly close. Then iterations will be done using the simulator. The gate inductance L_G is added to result in series resonance at 5.2 GHz from the CMOS equivalent of (7.70)

$$L_G = \frac{1}{C_{gs}\omega^2} \quad L_s = \frac{1}{230 \text{ fF} \times 2\pi \times 5.2 \text{ GHz}} \quad L_s = 4.07 \text{ nH} \quad 0.284 \text{ nH} = 3.79 \text{ nH}$$

The cascode transistor and the source follower buffer were both sized at 100 μm to reduce the parasitic capacitance and also to operate somewhat closer to the peak f_T current.

Through simulations, final spiral inductors for L_S and L_G were determined whose values at 5.2 GHz were 3.73 nH and 381 pH. Rather arbitrarily, a 1.1-nH inductor was chosen for the tank inductance requiring roughly a 0.85-pF capacitor to resonate at 5.2 GHz. In simulations, the capacitor value needed to be decreased to 700 fF to account for the roughly 150 fF of parasitic capacitance, in particular, that of the source follower transistor. All of the inductors had a Q of about 15; hence, the output parallel resistance can be calculated to be 539 . To reduce the gain somewhat, a 1.8k resistor was placed parallel to the tank circuit.

The gain to the drain can be predicted from (7.38). The output source follower further reduces the gain according to its output impedance, approximately equal to $1/g_m$ of the output transistor. The buffer current was adjusted to result in approximately 50 output impedance, although if the LNA were going to other on-chip components (e.g., a mixer), there would not necessarily be the need for a 50 output impedance. The predicted overall voltage gain is

$$\left|\frac{v_{\text{out}}}{v_{\text{in}}}\right| = \frac{R_L \omega_T}{R_S \omega_o} \frac{50}{r_o + 50} = \frac{539 \| 1{,}800}{50} \frac{14}{5.2} \frac{50}{100} = 11.17 \quad 21 \text{ dB}$$

In the derivation of gain, C_{gd} was ignored resulting in a simple expression containing the ratio of g_m to C_{gs}, which was replaced with the term ω_T. However, C_{GD} can be a significant fraction of C_{GS}, and even with a cascode structure there is a signal on the drain, meaning there is a Miller multiplication of this capacitor. Hence, effective gain is typically significantly decreased compared to the simple equation. In the above derivation, this was taken into account by using a value of 14 GHz for f_T about half of what is measured for the single transistor. The actual simulated gain shown in Figure 7.24(a) was 18.7 dB for a bias current of 2 mA, somewhat of a reduction due to additional parasitics not included above, for example, the loss in the gate and source inductances. Figure 7.24(b) shows that the input impedance is matched to better than 15 dB for any bias currents of at least 2 mA.

With a Q of about 15, the 3.73-nH gate inductor has a series resistance r_{LG} of about 8.12 at 5.2 GHz. The noise figure of a CS amplifier can be estimated from (7.66)

$$F = 1 + \frac{r_g + r_{LG}}{R_s} + \frac{\gamma}{g_m R_s} + \frac{\delta g_m R_s}{5} \left(\frac{\omega}{\omega_T}\right)^2$$

$$= 1 + \frac{8.12}{50} + \frac{0.67}{0.04 \times 50} + \frac{1.33 \times 0.04 \times 50}{5} \left(\frac{5.2}{28}\right)^2$$

$$F = 1 + 0.162 + 0.333 + 0.018 = 1.513 \quad 1.80 \text{ dB}$$

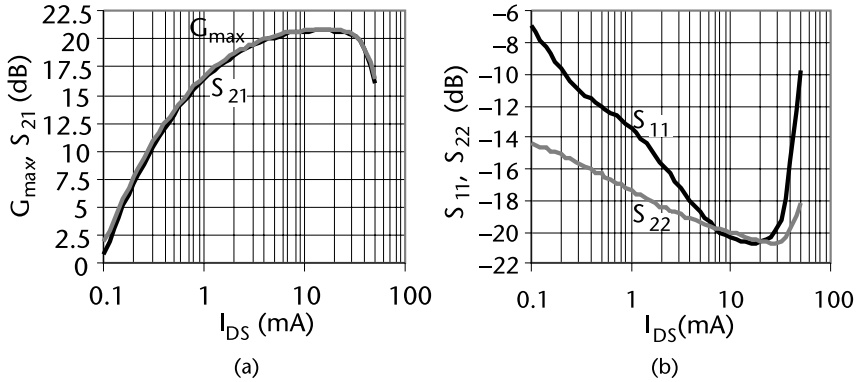

Figure 7.24 Transistor performance versus bias current: (a) G_{max} and S_{21} and (b) input matching (S_{21}) and output matching (S_{22}).

Since the transistors are operated at very low current densities, the long channel approximations are used for γ and δ, as discussed in Chapter 4, that is, $\gamma = 0.67$ and $\delta = 1.33$.

It can be noted that the above equation for the noise factor is valid for a common-source amplifier. However, this amplifier has matching components L_S and L_G and this affects the noise somewhat. In particular, since gain is no longer given by g_m multiplied by R_L but is determined by (7.38), this must be used to convert the drain channel noise into the equivalent input noise. As well, because of the gate inductance L_G and because of impedance matching, the noise current into the gate node sees an impedance that can be shown to be approximately $(1 + Q^2)R_s$ where Q is given by $\omega L_G/R_S$. The resulting equation for F and the resulting numerical value is

$$F = 1 + \frac{r_g + r_{LG}}{R_s} + \gamma g_m \times 4R_s \left(\frac{\omega}{\omega_T} \div\right)^2 + \frac{\delta g_m R_s (1 + Q^2)}{5} \left(\frac{\omega}{\omega_T} \div\right)^2$$

$$F = 1 + \frac{8.12}{50} + 0.67 \times 0.04 \times 4 \times 50 \left(\frac{5.2}{28} \div\right)^2 + \frac{1.33 \times 0.04 \times 50(1 + 2.44^2)}{5} \left(\frac{5.2}{28} \div\right)^2$$

$$F = 1 + 0.162 + 0.183 + 0.128 = 1.473 \quad 1.68 \text{ dB}$$

Simulated noise figure, shown in Figure 7.25, is 1.6 dB, close to the predicted value. By examining the simulated transistor, small-signal parameters, and noise summary, it seems that the transistor is approaching subthreshold operation; thus, even smaller values for γ and δ could be used. It is also noted that when designing amplifiers with very low noise, even small noise contributions, for example, from the bias resistor and from the cascode transistor, can end up having a significant impact on noise.

7.3.5 Relationship Between Noise Figure and Bias Current

Noise due to the base resistance is in series with the input voltage so it sees the full amplifier gain. The output noise due to base resistance is given by:

$$v_{no,r_b} \quad \sqrt{4kTr_b} \quad g_{m1}R_L \tag{7.71}$$

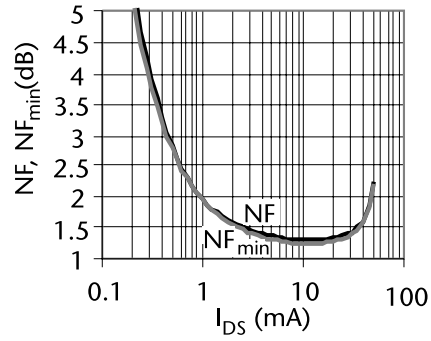

Figure 7.25 Noise figure and NF_{min} versus bias current.

Note that this noise voltage is proportional to the collector current, as is the signal, so the SNR is independent of bias current.

The collector shot noise is in parallel with the collector signal current and is directly sent to the output load resistor.

$$v_{no,I_c} \quad \sqrt{2qI_C}R_L \tag{7.72}$$

Note that this output noise voltage is proportional to the square root of the collector current, while the signal voltage is directly proportional to current; therefore, to improve the noise figure due to collector shot noise, we increase the current.

Base shot noise can be converted to input voltage by considering the impedance on the base. If Z_{eq} is the impedance on the base (formed by a combination of matching, base resistance, source resistance, and transistor input impedance), then:

$$v_{no,I_B} \quad \sqrt{\frac{2qI_C}{\beta_o}}Z_{eq}g_mR_L \tag{7.73}$$

Note that this output voltage is proportional to the collector current raised to the power of 3/2. Therefore, to improve the noise figure due to base shot noise, we decrease the current because the signal-to-noise ratio (in voltage terms) is inversely proportional to the square root of the collector current.

Therefore, at low currents collector shot noise will dominate and noise figure will improve with increasing current. However, the effect of base shot noise also increases and eventually will dominate. Thus, there will be some optimum level to which the collector current can be increased, beyond which the noise figure will start to degrade again. Note that this simple analysis ignores the fact that r_b increases at lower currents, so, in practice, thermal noise density due to r_b is more important at low currents than is indicated by this analysis.

For CMOS transistors, the effect on noise figure to a change of current is similar to that for the bipolar transistor, with some notable exceptions. For CMOS circuits, for a fixed load resistor, the signal power at the output for a constant input power is related to the square of g_m; hence, with the simple square law model, the output signal power is proportional to the current

$$v_o^2 = v_i^2 \quad g_{m1}^2R_L^2 \quad \alpha \quad I \tag{7.74}$$

In comparison, for bipolar devices, the output power increases with the square of current since g_m is proportional to the current. For noise due to gate resistance (or other series resistance), the gain to the output is the same as for the signal,

$$v_{no,r_g}^2 \quad 4kTr_g \quad g_{m1}^2R_L^2 \quad \alpha \quad I \tag{7.75}$$

hence, the noise figure is not affected by a change of current. This is the same as for bipolar circuits. For drain channel noise, output noise power is proportional to g_m

$$v_{no,i_d}^2 \quad 4kT\gamma g_m \quad R_L^2 \quad \alpha \quad \sqrt{I} \tag{7.76}$$

hence, to the square root of current. However, since the signal power is increasing in proportion to the current, the noise figure is improved with increasing current,

similar to the bipolar case. Finally, the gate induced noise power is inversely proportional to g_m,

$$v_{no,i_g}^2 \quad \frac{4kT\delta\omega^2 C_{gs}{}^2}{5} g_m \quad R_L^2 \quad \alpha \quad \sqrt{I} \tag{7.77}$$

hence, the noise figure will also improve with the increasing current, unlike the noise figure due to the base shot noise in a bipolar circuit, which degrades with an increasing current. Thus, with CMOS circuits with the simple models, the noise figure generally improves with an increasing bias current, and an optimum point is hard to identify. However, with the full models as used in a simulator, an optimum point will actually be seen due to other effects beyond the simple models. For example, as mentioned earlier, the effective value for γ used to determine the drain channel noise increases with bias current. As well, since the signal does not keep increasing with increasing current, for example, when the current density approaches or exceeds that for the maximum $f_{\max}$, the noise figure will degrade.

7.3.6 Effect of the Cascode on Noise Figure

As discussed in Section 7.1.3, the cascode transistor is a common-base amplifier with a current gain close to 1. By *Kirchoff's current law* (KCL) of the dotted box in Figure 7.26, $i_{c2} = i_{e2} - i_{b2} \quad i_{e2}$. Thus, the cascode transistor is forced to pass the current of the driver on to the output. This includes signal and noise current. Thus, to a first order, the cascode can have no effect on the noise figure of the amplifier. However, in reality it will add some noise to the system. For this reason, the cascode LNA can never be as low noise as a common-emitter amplifier.

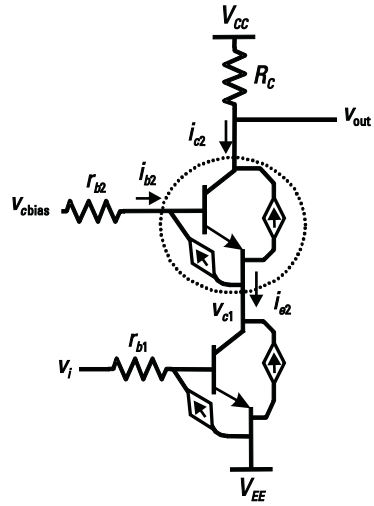

Figure 7.26 A cascode LNA showing noise sources.

7.3.7 Noise in the Common-Collector/Drain Amplifier

Since this type of amplifier is not often used as an LNA stage, but more commonly as a buffer, we will deal with its noise only briefly. The amplifier with noise sources is shown in Figure 7.27. Noise due to r_b is directly in series with noise due to R_S. If noise at the output were due only to R_S, the noise figure would be 0 dB.

The noise due to collector shot noise is reduced due to negative feedback caused by R_E. For example, the current added causes the v_{be} to decrease (as the v_e increases), and i_e is decreased, counteracting the added current. Note that noise due to R_E sees the same effect (negative feedback reduction of noise).

The base shot noise current is injected into the base where an input voltage is developed across R_S: v_{nbs} $i_{nbs}R_S$ v_o. The exact relationship between the output noise voltage v_{o_nbs} due to base shot noise current i_{nbs} is:

$$v_{o\_nbs} = \frac{R_E Z_\pi (1 \quad g_m R_S)}{R_E(1 + g_m Z_\pi) + R_S + r_\pi} i_{nbs} \tag{7.78}$$

Assuming that R_E is large, $g_m R_S >> 1$, $g_m Z_\pi >> 1$, and then:

$$v_{o\_nbs} = \frac{R_E Z_\pi (1 \quad g_m R_S)}{R_E(1 + g_m Z_\pi)} i_{nbs} \quad \frac{R_E Z_\pi g_m R_S}{R_E g_m Z_\pi} i_{nbs} \quad R_S i_{nbs} \tag{7.79}$$

The relationship between the collector shot noise i_{ncs} and the output noise voltage v_{o_ncs} can be shown to be:

$$v_{o\_ncs} = \frac{R_E(R_S + Z_\pi)}{R_E(1 + g_m Z_\pi) + Z_\pi + R_S} i_{ncs} \tag{7.80}$$

Assuming that R_E is large, and $R_S >> Z_\pi$, then:

$$v_{o\_ncs} \quad \frac{R_S + Z_\pi}{1 + g_m Z_\pi} i_{ncs} \quad r_e i_{ncs} \tag{7.81}$$

Therefore, the collector shot noise current sees r_e, a low value, and output voltage is low. Thus, the common-collector adds little noise to the signal except through r_b.

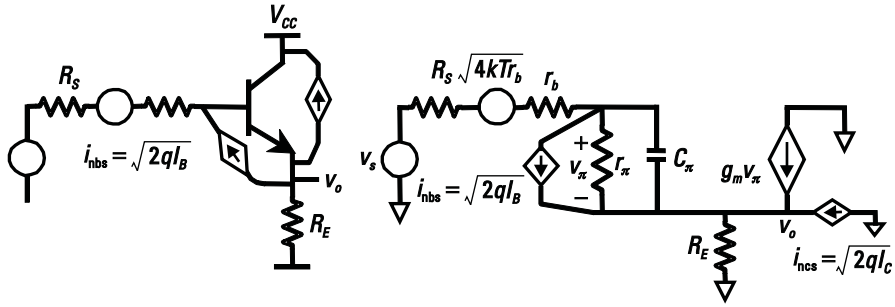

Figure 7.27 A common-collector amplifier with noise illustrated.

7.4 Linearity in Amplifiers

Nonlinearity analysis will follow the same basic principles as discussed in Chapter 2, with power series expansions and nonlinear terms present in the amplifier. These will now be discussed in detail.

7.4.1 Exponential Nonlinearity in the Bipolar Transistor

In bipolar transistors, one of the most important nonlinearities present is the basic exponential characteristic of the transistor itself, illustrated in Figure 7.28.

Source resistance improves linearity. As an extreme example, if the input is a current source, $R_S =$, then $i_c = \beta i_b$. This is as linear as β is. It can be shown that a resistor in the emitter of value R_E has the same effect as a source or base resistor of value $R_E\beta$. The transistor base has a bias applied to it and an ac signal superimposed. Summing the voltages from ground to the base and assuming that $i_e = i_c$:

$$v_s + V_S = v_{be} + V_{BE} + R_E(I_C + i_c) \tag{7.82}$$

where V_{BE} and v_{be} are the dc and ac voltages across the base emitter junction of the transistor.

Extracting only the small signal components from this equation gives:

$$v_s = v_{be} + R_E i_c \tag{7.83}$$

Also from the basic properties of the junction:

$$I_c + i_c = I_S e^{\frac{V_{BE}+v_{be}}{v_T}} = I_S e^{\frac{V_{BE}}{v_T}} e^{\frac{v_{be}}{v_T}} = I_C e^{\frac{v_{be}}{v_T}} \tag{7.84}$$

where from Chapter 4, $v_T = kT/q$. Solving for v_{be} gives:

$$v_{be} = v_T \ln\left[1 + \frac{i_c}{I_C}\right] \tag{7.85}$$

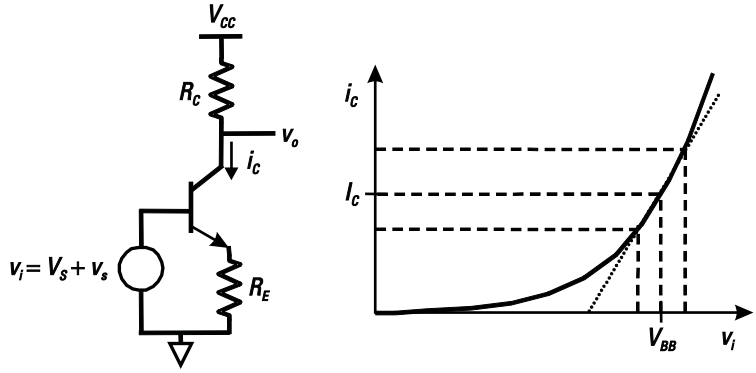

Figure 7.28 Bipolar common-emitter amplifier for linearity analysis.

Now making use of the math identity:

$$\ln(1+x) = x \quad \frac{1}{2}x^2 + \frac{1}{3}x^3 \ldots \tag{7.86}$$

and expanding (7.85) using (7.86) and substituting it back into (7.83), we get:

$$v_s = R_E i_c + v_T \quad \frac{i_c}{I_C} \quad \frac{1}{2} \quad \frac{i_c}{I_C} \div \quad ^2 + \frac{1}{3} \quad \frac{i_c}{I_C} \div \quad ^3 \ldots \tag{7.87}$$

Noting that $v_T/I_C = r_e$ and rearranging, we get:

$$v_s = (\mathrm{R}_E + r_e) i_c \quad \frac{1}{2} r_e \frac{i_c^{\,2}}{I_C} + \frac{1}{3} r_e \frac{i_c^{\,3}}{I_C^{\,2}} \ldots \tag{7.88}$$

This can be further manipulated to give:

$$\frac{v_s}{(R_E + r_e)} = i_c \quad \frac{1}{2} \frac{r_e}{(R_E + r_e)} \frac{i_c^{\,2}}{I_C} + \frac{1}{3} \frac{r_e}{(R_E + r_e)} \frac{i_c^{\,3}}{I_C^{\,2}} \ldots \tag{7.89}$$

This is the equation we need, but it is in the wrong form. It needs to be solved for i_c. Thus, a few more relationships are needed. Given:

$$y = a_1 x + a_2 x^2 + a_3 x^3 + \ldots \tag{7.90}$$

the following can be found:

$$x = b_1 y + b_2 y^2 + b_3 y^3 + \ldots \tag{7.91}$$

where

$$\begin{aligned} b_1 &= \frac{1}{a_1} \\ b_2 &= \quad \frac{a_2}{a_1^{\,3}} \\ b_3 &= \frac{1}{a_1^{\,5}} (2a_2^{\,2} \quad a_1 a_3) \end{aligned} \tag{7.92}$$

Equation (7.89) can now be rewritten as a function of i_c:

$$\begin{aligned} i_c = \frac{v_s}{R_E + r_e} + \frac{1}{2I_c} \quad \frac{r_e}{R_E + r_e} \div \quad \frac{v_s}{R_E + r_e} \div \quad ^2 \\ + \quad \frac{1}{2I_c^{\,2}} \quad \frac{r_e}{R_E + r_e} \div \quad ^2 \quad \frac{1}{3I_c^{\,2}} \quad \frac{r_e}{R_E + r_e} \div \quad \frac{v_s}{R_E + r_e} \div \quad ^3 \end{aligned} \tag{7.93}$$

Now the third-order intercept voltage can be determined (note that this is the peak voltage and the rms voltage will be lower by a factor of $\sqrt{2}$):

$$v_{IP3} = 2\sqrt{\frac{k_1}{3k_3}} = 2\sqrt{\frac{1}{3}\ \frac{1}{|R_E + r_e|} \div \frac{6I_C{}^2\left(|R_E + r_e|\right)^5}{r_e|2R_E \quad r_e|}} = 2\sqrt{2}v_T \frac{|R_E + r_e|^2}{\sqrt{r_e{}^3|2R_E \quad r_e|}} \quad (7.94)$$

This very useful equation can be used to estimate the linearity of gain stages. An approximation to (7.94) that can be quite useful for hand calculations is:

$$v_{IP3} = 2\sqrt{2}v_T \left(\frac{R_E + r_e}{r_e}\right)^{\frac{3}{2}} \quad (7.95)$$

In the special case where there is no emitter degeneration, the above expression can be simplified to:

$$v_{IP3} = 2\sqrt{2}v_T \quad (7.96)$$

Example 7.11: Linearity Calculations in Common-Emitter Amplifier

For a common-emitter amplifier with no degeneration, if the input is assumed to be composed of two sine waves of amplitude A_1 and A_2, compute the relevant frequency components to graph the fundamental and third-order products and predict what the IIP3 point will be. Assume that $I_{CA} = 1$ mA and $v_T = 25$ mV.

Solution:
The first step is to calculate the coefficient k_1, k_2, and k_3 for the power series expansion from (7.93) as:

$$k_1 = \frac{1}{R_E + r_e} = \frac{1}{0 + 25} = 0.04$$

$$k_2 = \frac{1}{2I_C}\left(\frac{r_e}{R_E + r_e} \div \frac{1}{R_E + r_e} \div\right)^2 = \frac{1}{2I_C}\frac{r_e}{(R_E + r_e)^3} = \frac{1}{2 \quad 1\text{m}}\frac{25}{(0 + 25)^3} = 0.8$$

$$k_3 = \frac{1}{2I_c{}^2}\left(\frac{r_e}{R_E + r_e} \div\right)^2 \frac{1}{3I_c{}^2}\left(\frac{r_e}{R_E + r_e} \div\right)\left(\frac{1}{R_E + r_e} \div\right)^3$$

$$= 3\left(\frac{r_e}{R_E + r_e} \div\right)^2 2\left(\frac{r_e}{R_E + r_e} \div\right)\left(\frac{1}{6I_c{}^2} \div\right)\left(\frac{1}{R_E + r_e} \div\right)^3$$

$$= 3\left(\frac{25}{0 + 25} \div\right)^2 2\left(\frac{25}{0 + 25} \div\right)\left(\frac{1}{6 \quad 1\,\text{m}^2} \div\right)\left(\frac{1}{0 + 25} \div\right)^3$$

$$= [3 \quad 2]\left(\frac{1}{6 \quad 1\,\text{m}^2} \div\right)\left(\frac{1}{0 + 25} \div\right)^3 = 10.667$$

resulting in an expression for current as follows:

Table 7.1 Harmonic Components

Component	With A_1, A_2	With $(A_1 = A_2 = A)$	
dc	$k_o + \frac{k_2}{2}(A_1^2 + A_2^2)$	$k_o + k_2 A^2$	$1m + 0.8A^2$
Fundamental	$k_1 A_1 + k_3 A_1 \ \frac{3}{4}A_1^2 + \frac{3}{2}A_2^2 \div$	$k_1 A + \frac{9}{4}k_3 A^3$	$0.04A + 24A^3$
Second harmonic	$\frac{K_2 A_1^2}{2}$	$\frac{K_2 A^2}{2}$	$0.4A^2$
Intermod	$\frac{3}{4}k_3 A_1^2 A_2$	$\frac{3}{4}k_3 A^3$	$8A^3$

$$i_c = k_1 v_s + k_2 v_s^2 + k_3 v_s^3 + \ldots = 0.04 v_s + 0.8 v_s^2 + 10.667 v_s^3 + \ldots$$

The dc component, fundamental, second harmonic, intermodulation components, and equations are listed for the above coefficients in Table 7.1.

The intercept point is at a voltage of 70.7 mV at the input and 2.828 mA at the output, as shown in Figure 7.29, which agrees with (7.96). For an input of 70.7 mV, the actual output fundamental current is 11.3 mA, which illustrates the gain expansion for an exponential nonlinearity.

The voltage-versus-current transfer function is shown in Figure 7.30. Also shown are the time-domain input and output waveforms demonstrating the expansion offered by the exponential nonlinearity.

Figure 7.30 illustrates a number of points about nonlinearity.

Due to the second-order term, k_2, there is a dc shift. Using the dc component term as shown in Table 7.1, we make the following calculations.

With $A = 70.71$ mV, the shift is:

$$0.8 \quad (70.71\,\text{mV})^2 = 4.0\,\text{mA}$$

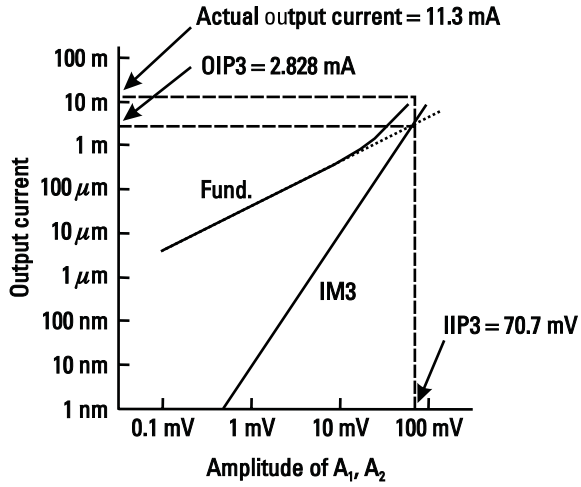

Figure 7.29 Plot of fundamental and third-order products coming out of an exponential nonlinearity.

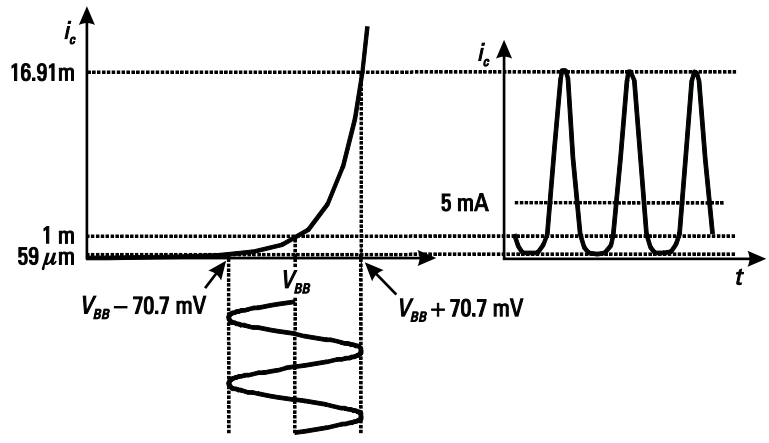

Figure 7.30 Input output transfer function and time-domain voltage and current waveforms.

Thus, the waveform is centered on 5 mA.

Because of positive coefficient k_3, the waveform is not compressed, but expanded. However, either way, compression or expansion, the result is distortion.

The above calculations all assume that R_C is small enough so that the transistor does not saturate. If saturation does occur, the power series is no longer valid.

Typically, inputs would not be allowed to be bigger than about 10 dB below IP3, which, for this example, is about 22.34 mV. Figure 7.31 shows the transfer function for an input of this amplitude. At this level, the current goes from 0.409 to 2.444 mA. The dc shift is 0.4 mA, so the current is 1.4 ± 1.01 mA.

Example 7.12: Linearity Calculations in Common-Emitter Amplifier with Degeneration

Continue the previous example by determining the effect of emitter degeneration. For an input of two sine waves of amplitude A_1 and A_2, plot IIP3 for R_E ranging from 0 to 50 . Again, assume that $I_{CA} = 1$ mA, and $v_T = 25$ mV.

Solution:

To determine IP3, two tones can be applied at various amplitudes and graphical extrapolations made of the fundamental and third-order tones, as previously

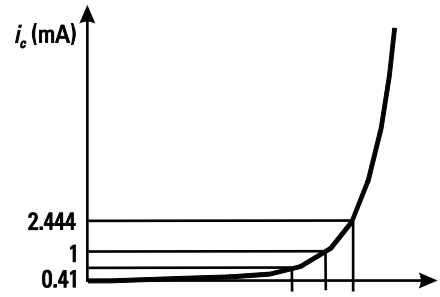

Figure 7.31 Transistor characteristic for smaller input signal.

illustrated in Figure 7.29. Instead, for this example, values for k_0, k_1, k_2, k_3 are determined from the equations in Table 7.1. Then these are used to calculate the fundamental and *third-order intermodulation* (IM3) components from the equations in Table 7.1 and IIP3 calculated from the given fundamental and third-order terms similar to that discussed in Section 2.3.2.

$$\text{IIP3} = 10\log\frac{A^2}{2 \quad 50 \quad 1\ \text{mW}}\div + \frac{1}{2} \quad 20\log\frac{\text{Fundamental}}{|\text{IM3}|}\div$$

$$= \quad 50 + 10\log\frac{\text{Fundamental}}{|\text{IM3}|}\div$$

Figure 7.32 shows the resulting plot of IIP3 and, for comparison, the approximation of (7.95).

This example clearly shows how IP3 improves with degeneration resistance. From the fundamental equations, it can also be seen that for larger I_C and hence larger I_B, the improvement will be higher. It can also be seen from the equation that it is possible to cancel the third-order term if $R_E = r_e/2$, which in this example requires a degeneration of 12.5 . It can be seen that for lower degeneration, k_3 is positive, resulting in gain expansion, while for larger values of degeneration, k_3 is negative, resulting in gain compression. At exactly 12.5 , k_3 goes through zero and theoretical IP3 goes to infinity. In real life, if k_3 is zero, there will be a component from the k_5 term, which will limit the linearity. However, this improvement in linearity is real and can be demonstrated experimentally [3].

A related note of interest is that for a MOSFET transistor operated in subthreshold, the transistor drain characteristics are exponential and hence k_3 is positive, while for higher bias levels, k_3 is negative. Thus, by an appropriate choice of bias

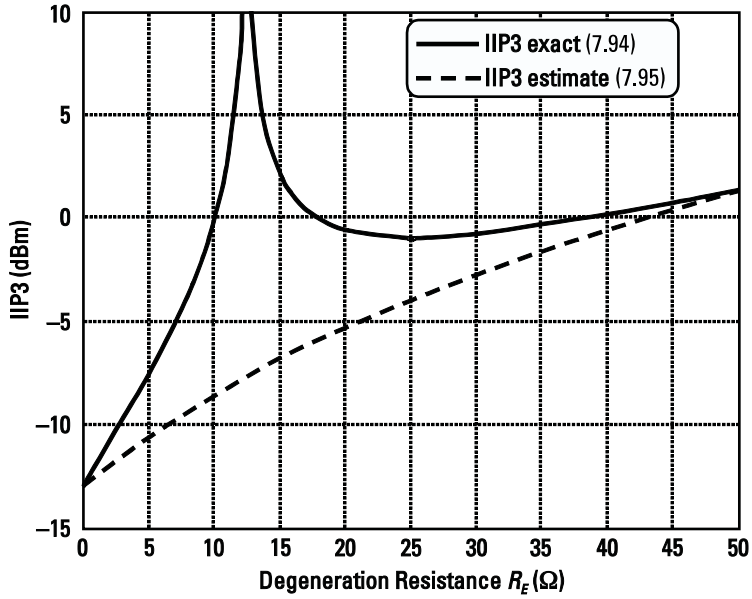

Figure 7.32 Input IP3 as a function of degeneration resistance.

conditions, k_3 can be set to zero for improvements in linearity [4]. In MOSFETs, it turns out to be quite challenging to take advantage of this linearity improvement, since the peak occurs for a narrow region of bias conditions and the use of degeneration resistance or inductance reduces this linearity improvement.

7.4.2 Nonlinearity in the CMOS Transistor

With perfect square law equations, it would seem there should be no third-order intermodulation. However, there are several reasons why third-order nonlinearity occurs. The standard square law equations are always accompanied with additional constraints that the input voltage is bigger than the threshold voltages ($v_{GS} > v_T$) and that the drain-source voltage is larger than the on voltage ($v_{DS} > v_{GS} - v_T$). With large input or output voltages, these limits will introduce clipping and hence a third-order component. The small signal output current is related to the input by the transconductance g_m, where g_m is the derivative of current with respect to the input voltage. The voltage required to cause clipping can be estimated as I_o/g_m. This voltage is roughly equal to the one-tone 1-dB compression voltage, and v_{IP3} can be estimated to be larger by roughly 9.6 dB, or a voltage ratio of 3.02. For square law equations as given in Chapter 4, this can be expressed as:

$$v_{IP3} \quad 3.02\frac{I_o}{g_m} = 3.02\frac{\frac{\mu C_{ox}}{2}\frac{W}{L}(v_{GS} \quad v_T)^2(1+\lambda v_{DS})}{\mu C_{ox}\frac{W}{L}(v_{GS} \quad v_T)(1+\lambda v_{DS})} = 1.51(v_{GS} \quad v_T) \quad (7.97)$$

Thus, given an operating current I_o, linearity can be estimated if the transconductance is known or if the input voltage and threshold voltage are known.

Furthermore, all real CMOS transistors do not follow the square law model; hence, there will be some third-order component even without clipping. Finally, by adding degeneration, a third-order term is introduced due to the second-order term feeding back to the input and mixing with the fundamental component.

7.4.3 Nonlinearity in the Output Impedance of the Bipolar Transistor

Another important nonlinearity in the bipolar or CMOS transistor is the output impedance. An example of where this may be important is in the case of a transistor being used as a current source. In this circuit, the base of the transistor is biased with a constant voltage and the current into the collector is intended to remain constant for any output voltage. Of course, the transistor has a finite output impedance, so if there is an ac voltage on the output, there is some finite ac current that flows through the transistor, as shown by r_o in Figure 7.33. Worse than this, however, is the fact that the transistor's output impedance will change with applied voltage and therefore can introduce nonlinearity.

The dc output impedance of a transistor is given by:

$$r_{o\_DC} = \frac{V_A}{I_C} \quad (7.98)$$

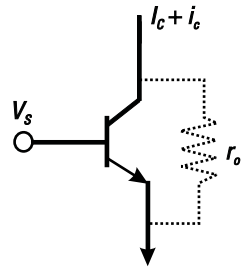

Figure 7.33 Bipolar transistor as a current source.

where V_A is the Early voltage of the transistors. An ac current into the collector can be written as a function of ac current i_c.

$$r_{o\_ac}(i_c) = \frac{V_A}{I_C + i_c} \tag{7.99}$$

Assuming for this analysis, that there is no other significant impedance in the circuit other than the transistor output resistance, the ac collector-emitter voltage can be written as:

$$v_{ce} = i_c r_{o\_ac} = \frac{i_c V_A}{I_C + i_c} = V_A \frac{\frac{i_c}{I_C}}{1 + \frac{i_c}{I_C}} \tag{7.100}$$

Now from the relationship:

$$\frac{x}{1+x} = x \quad x^2 + x^3 \quad x^4 + x^5 \ldots \tag{7.101}$$

(7.100) can be written out as a power series:

$$v_{ce} = V_A \left(\frac{i_c}{I_C} \div \right) \quad V_A \left(\frac{i_c}{I_C} \div \right)^2 + V_A \left(\frac{i_c}{I_C} \div \right)^3 = r_{o\_DC} i_c \quad \frac{r_{o\_DC}}{I_C} {i_c}^2 + \frac{r_{o\_DC}}{{I_C}^2} {i_c}^3 \tag{7.102}$$

The intermodulation current can now be easily determined.

$$i_{IP3} = 2\sqrt{\frac{k_1}{3k_3}} = 2\sqrt{\frac{1}{3} r_{o\_DC} \frac{{I_C}^2}{r_{o\_DC}}} = \frac{2I_C}{\sqrt{3}} \tag{7.103}$$

Thus, the output intermodulation voltage is just:

$$v_{OP3} = \frac{2I_C}{\sqrt{3}} r_o \tag{7.104}$$

This is a fairly intuitive result. As the dc current is increased, the ac current is a smaller percentage of the total and therefore the circuit behaves more linearly. Thus,

the designer has two choices if the current source is not linear enough. They can either increase the current or increase the output impedance. Also, it should be noted that this relationship only holds true if the transistor does not start to saturate. If it does, the nonlinearity will get much worse.

7.4.4 High-Frequency Nonlinearity in the Bipolar Transistor

Many frequency-dependent devices can reduce the linearity of a circuit. One of the most troublesome is the base-collector junction capacitance C_μ. This capacitance is voltage dependent, which results in a nonlinearity. This nonlinearity is especially important in circuits with low supply voltages because the capacitance is largest at a low reverse bias.

This capacitor's effect is particularly harmful for both frequency response and nonlinearity in the case of a standard common-emitter amplifier. In this configuration, C_μ is multiplied by the gain of the amplifier (the Miller effect) and appears across the source.

The value of C_μ as a function of bias voltage is given by:

$$C_\mu(V) = \frac{C_{\mu o}}{1 \quad \frac{V}{\psi_o} \div^{(1/n)}} \tag{7.105}$$

where $C_{\mu o}$ is the capacitance of the junction under zero bias, ψ_o is the built-in potential of the junction, and n is usually between 2 and 5. Since this capacitor's behavior is highly process-dependent and hard to model, there is little benefit in deriving detailed equations for it. Rather, the designer must rely on simulation and detailed models to predict its behavior accurately.

7.4.5 Linearity in Common-Collector/Drain Configuration

The common-collector amplifier is often called the *emitter-follower* because the emitter voltage "follows" the base voltage. However, the amplifier cannot do this over all conditions. If the current is constant, v_{BE} is constant and the transfer function will be perfectly linear. However, as v_o changes, $i_{\text{out}} = v_o/R_{\text{out}}$ will change as shown in Figure 7.34. Thus, i_E will change, and so will v_{BE} and there will be some nonlinearity.

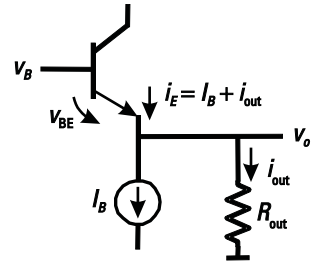

Figure 7.34 Illustration of nonlinearity in the common-collector amplifier.

If R_{out} is large enough that i_{out} is always much less than I_B, the linearity will be good, as the operating point will not change significantly over a cycle of the signal. It is important to keep the peak output current less than the bias current. This means that:

$$\frac{|v_{o,peak}|}{R_{out}} < I_B \tag{7.106}$$

If this is the case, then there will be no clipping of the waveform.

The linearity can be improved by increasing I_B, or R_{out}. This will continue to improve performance as long as the power supply voltage is large enough to allow this swing. Thus, for large R_{out}, the power supply limits the voltage swing and therefore the linearity. In this case, the current is not a limiting factor.

7.5 Stability

A number of different but equivalent measures of stability can be derived from S-parameters [5]. For RFIC design, the stability factor K and the auxiliary stability factor B_1 are typically determined directly using a simulator rather than calculating them from the S-parameters. Unconditional stability is guaranteed by the following two conditions:

$$\begin{aligned} K &> 1 \\ B_1 &> 0 \end{aligned} \tag{7.107}$$

For the sake of completeness, these are based on S-parameters as follows [5]:

$$K = \frac{1 \quad |S_{11}|^2 \quad |S_{22}|^2 + |\quad|^2}{2\,|S_{12}S_{21}|} \tag{7.108}$$

and

$$B_1 = 1 + |S_{11}|^2 \quad |S_{22}|^2 \quad |\quad|^2 \tag{7.109}$$

where

$$= S_{11}S_{22} \quad S_{12}S_{21} \tag{7.110}$$

Instability at any frequency, not just at the operating frequency, can result in unwanted oscillations that render the circuit inoperable and hence must be avoided. Thus, frequencies are swept and K and B_1 are determined over the full frequency range for which the transistor has gain. For any frequency for which K is less than 1 or B_1 is less than 0, the circuit is not unconditionally stable, meaning that a load or source impedance exists for which the circuit will be unstable. Having identified problem frequencies, a typical further step might be to plot source and load stability circles at the problem frequency to determine the impedance for which the instability occurs. If such problem impedances are safely away from the actual impedance, it may not be necessary to modify the circuit. However, in some cases the source or

load impedance is not fully known for all frequencies, and hence making the circuit unconditionally stable would be recommended. For example, for a power amplifier or LNA connected to a 50 antenna, it can happen that a potential stability problem is identified away from the operating frequency, but only for an impedance away from 50 . Even though the antenna may be 50 at the operating frequency, instability is a real possibility since the antenna impedance at the other frequencies is expected to be quite different.

Techniques to improve stability will depend on the frequency, impedance, and location of stability problems as identified by the stability factors over frequency, input and output stability circles, observation of time-domain oscillations, and discontinuities in the maximum gain plots. Techniques to improve stability can include adding small series resistance or large parallel resistance at the input or output. We note that real series inductors with finite Q can improve stability because of their series resistance, while the use of ideal inductors can make stability problems worse. Also, frequency selective networks may work if the problem frequencies are away from the operating frequency. For example, highpass or lowpass matching networks may be used. The use of a cascode structure can improve reverse isolation (feedback) and this can help with stability; however, it is important to have the base of the cascode transistor (or the gate of a CMOS cascode transistor) at low impedance, for example, bypassed with a decoupling capacitor; otherwise, the cascode structure can degrade stability.

7.6 Differential Amplifiers

Any of the amplifiers that have already been discussed can be made differential by adding a mirrored copy of the original and connecting them together at the points of symmetry so that voltages are no longer referenced to ground, but rather swing plus or minus relative to each other. This will be illustrated in the following sections. The advantage of differential amplifiers and differential circuits in general is that the positive and negative signals are referenced to each other instead of to ground. As a result, the connection to the off-chip ground is less critical and as a result packaging is less important. As well, noise coupled from the substrate or from external sources often appears roughly equally on positive and negative nodes; hence, for signals that are processed differentially, such noise is reduced. Finally, the signal swing can be increased in differential circuits since either side can swing positive and negative from the nominal operating point. On the negative side, differential circuits typically require more transistors and more current as there are two signal paths. As well, since the positive signal is referenced only to the negative signal, in a high-gain circuit additional circuits called common-mode feedback circuits are used to ensure that the operating point or average value of the signals is correct.

7.6.1 Bipolar Differential Pair

While this is hard to describe, it is easy to show an example of a differential common-emitter amplifier (more commonly called a differential pair or emitter-coupled pair)

in Figure 7.35. Here the bias for the stage is supplied with a current source in the emitter. Note that when the bias is applied this way, the emitter is at a virtual ground. This means that for small-signal differential inputs, this voltage never moves from its nominal voltage.

This stage can be used in many circuits such as mixers, oscillators, or dividers. If an input voltage is applied larger than about $5v_T$, then the transistors will be fully switched and they can act as a limiting stage or "square wave generator" as well.

An equation for current can be found by expressing i_{C1} and i_{C2} as exponentials of the input voltage and summing the two currents to be equal to I_{EE}. After a few lines of algebra, the resulting equation, which can be written with an exponential or with a hyperbolic tangent, is shown in (7.111) and (7.112).

$$i_{C1} = \frac{I_{EE}}{1 + e^{(v_1/v_T)}} = \frac{I_{EE}}{2}\left[1 + \tanh\left(\frac{v_1}{2v_T}\right)\right] \tag{7.111}$$

$$i_{C2} = \frac{I_{EE}}{1 + e^{(v_1/v_T)}} = \frac{I_{EE}}{2}\left[1 - \tanh\left(\frac{v_1}{2v_T}\right)\right] \tag{7.112}$$

and the differential output voltage is given by:

$$v_{o2} - v_{o1} = I_{EE}R_C \tanh\left(\frac{v_1}{2v_T}\right) \tag{7.113}$$

Note that if parasitic capacitance can be ignored, there will only be odd order terms in a power series expansion of this nonlinearity and hence no dc shifts or even-order harmonics in $v_{o2} - v_{o1}$ as v_1 grows. At higher frequencies where parasitic capacitance becomes increasingly important, second-order terms will become correspondingly important.

The slope of i_{C2} with respect to v_1 at $v_1 = 0$ will be:

$$\frac{i_{C2}}{v_1} = \frac{i_{C2}}{2}\frac{}{v_{BE2}} = \frac{1}{2}g_{m2} = \frac{1}{2}\frac{I_{C2}}{v_T} = \frac{I_{EE}}{4v_T} \tag{7.114}$$

This can be found directly by taking the derivative of the above equation for i_{C2} and setting v_1 to 0. We note that the linear extension of the small-signal slope

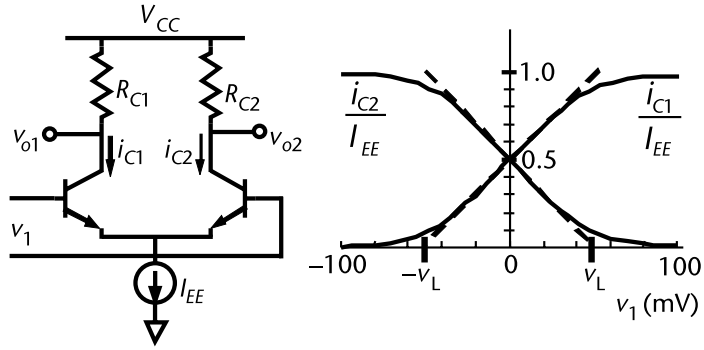

Figure 7.35 Differential common-emitter amplifier or emitter coupled pair.

intersects the axis at v_L given by the ratio of the nominal bias current I_C (equal to $I_{EE}/2$) to the effective transconductance g_{meff} (equal to $I_{EE}/4v_T$), or

$$v_L = \frac{I_C}{g_{meff}} = \frac{I_{EE}}{g_{m2}} = I_{EE}\frac{2v_T}{I_{EE}} = 2v_T \quad 50\text{ mV} \tag{7.115}$$

7.6.2 Linearity in Bipolar Differential Pairs

As shown in the previous section, the equation for differential output current in a bipolar differential pair with a total current bias of I_o is given by

$$i_o = I_o \tanh\frac{v_1}{2v_T} \tag{7.116}$$

The hyperbolic tangent can be expanded with

$$\tanh x = x \quad \frac{1}{3}x^3 + \frac{2}{15}x^5 \quad \ldots \tag{7.117}$$

$$\tanh\frac{v}{2v_T} = \quad \frac{v}{2v_T} \div \quad \frac{1}{3} \quad \frac{v}{2v_T} \div \quad ^3 + \frac{2}{15} \quad \frac{v}{2v_T} \div \quad ^5 \quad \ldots \tag{7.118}$$

$$i_o = \frac{I_o}{2v_T}v \quad \frac{I_o}{24{v_T}^3}v^3 + \ldots \tag{7.119}$$

$$v_{\text{IP3}} = \sqrt{\frac{4k_1}{3|k_3|}} = \sqrt{\frac{\frac{2I_o}{v_T}}{3\frac{I_o}{24{v_T}^3}}} = 4v_T = 2v_L \tag{7.120}$$

Thus it can be seen that v_{IP3} is equal to $4v_T$ or $2v_L$. Thus, no matter what the transistor size or bias current, v_{IP3} is equal to about 100 mV. To get increased linearity, add resistive degeneration in series with the emitters and then use the new value for v_L to estimate the improved linearity.

$$v_{\text{IP3}} = 2v_L = 2\frac{I_C}{g_{meff}} = 2\frac{I_{EE}}{2} \quad \frac{4v_T}{I_{EE}} + R_E \div = 4v_T + I_{EE}R_E = 2\frac{I_{EE}}{2}(2r_e + R_E) \tag{7.121}$$

where R_E is the total resistance between the two emitters.

The input one-tone 1-dB compression voltage is ideally lower than the input v_{IP3} by 9.66 dB or a linear factor of 3.02. Thus, $v_{1\text{dB}}$ is equal to $0.662v_L$. Because the resistor provides linear feedback, for large values of R_E linearity is expected to be better than predicted above, and a common estimate is that v_L is equal to the 1-dB compression point, in which case v_{IP3} would be equal to 3.02 v_L.

7.6.3 CMOS Differential Pair

The transfer characteristics of a CMOS differential pair, shown in Figure 7.36, can also be determined using the simple square law model for the transistor.

Given that

$$v_1 = v_{GS1} \quad v_{GS2} \tag{7.122}$$

We can solve for v_1 as function and of i_{D1} and i_{D2} using the square law relationship

$$i_D = \frac{\mu C_{ox}}{2} \frac{W}{L} \div (v_{GS} \quad V_T)^2 \tag{7.123}$$

and noting that

$$i_{D1} + i_{D2} = I_o \tag{7.124}$$

The solution for v_1 can be found as

$$v_1 = \sqrt{\frac{2}{\mu C_{ox}} \frac{L}{W} \div} \left(\sqrt{i_{D1}} \quad \sqrt{i_{D2}} \right) \tag{7.125}$$

or for i_{D1} as a function of v_1:

$$\begin{aligned} i_{D1} = & \frac{I_o}{2} + \frac{v_1}{2} \sqrt{I_o \mu C_{ox} \frac{W}{L} \div} \sqrt{1 \quad \frac{(\mu C_{ox})}{4 I_o} \frac{W}{L} \div v_1^2} \div \\ & = \frac{I_o}{2} + \frac{g_m v_1}{2} \sqrt{1 \quad \frac{(\mu C_{ox})}{4 I_o} \frac{W}{L} \div v_1^2} \end{aligned} \tag{7.126}$$

where

$$g_m = \sqrt{\mu C_{ox} \frac{W}{L} I_o} = \sqrt{K I_o} \tag{7.127}$$

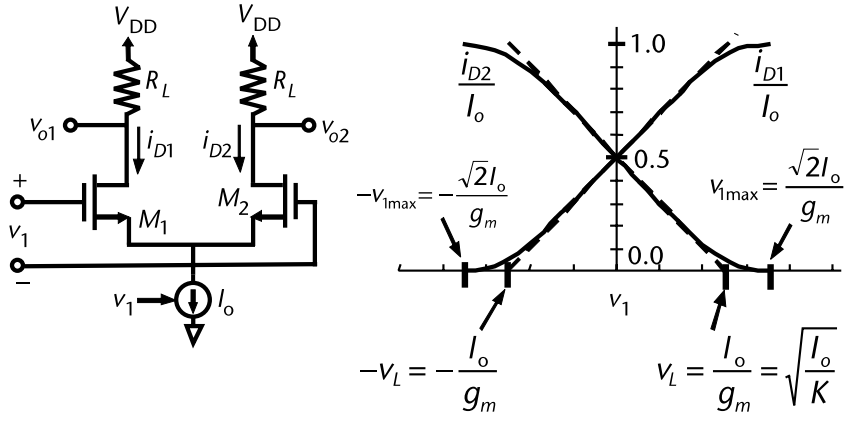

Figure 7.36 CMOS differential pair.

where the variable K gathers the process constants and transistor size as

$$K = \mu C_{ox} \left(\frac{W}{L} \right) \tag{7.128}$$

7.6.4 Linearity of the CMOS Differential Pair

The equation for current can be expressed as a power series by using the binomial expansion:

$$(1 \quad x)^{1/2} = 1 \quad \frac{x}{2} \quad \frac{1}{8}x^2 \quad \frac{1}{16}x^3 \ldots \tag{7.129}$$

where

$$x = \frac{(\mu C_{ox})}{4I_o} \left(\frac{W}{L} \right) v_2{}^2 = \frac{K}{4I_o} v_2{}^2 \tag{7.130}$$

As a result, i_{D1} can be expanded as

$$i_{D1} = \frac{I_o}{2} + \frac{v_2}{2}\sqrt{I_o K}(1 \quad x)^{1/2} \div = \frac{I_o}{2} + \frac{v_2}{2}\sqrt{I_o K} \quad 1 \quad \frac{K}{8I_o}v_2^2 \quad \frac{K^2}{128I_o^2}v_2^3 \quad \cdots \div \tag{7.131}$$

This simplifies to

$$i_{D1} = \frac{I_o}{2} + \frac{\sqrt{I_o K}}{2} v_2 \quad \frac{K}{16}\sqrt{\frac{K}{I_o}} v_2^3 \quad \ldots \tag{7.132}$$

where I_o is given by

$$I_o = 2I_{D1} = \mu C_{ox} \frac{W}{L} (v_{GS} \quad v_t)^2 \tag{7.133}$$

Finally, v_{IP3} can be expressed as

$$v_{IP3} = \sqrt{\frac{4k_1}{3|k_3|}} = \sqrt{\frac{4\frac{\sqrt{I_o K}}{2}}{3\frac{K}{16}\sqrt{\frac{K}{I_o}}}} = 4\sqrt{\frac{2}{3}\frac{I_o}{K}} = 4\sqrt{\frac{2}{3}\frac{I_o L}{\mu C_{ox} W}} = 4\sqrt{\frac{2}{3}}(v_{GS} \quad v_t) \tag{7.134}$$

or

$$v_{IP3} = 4\sqrt{\frac{2}{3}\frac{I_o}{K}} = 4\sqrt{\frac{2}{3}} v_L = 3.266 v_L \tag{7.135}$$

It can be noted that the factor of 3.226 is equivalent to 10.3 dB. This is close to the 9.66 dB that v_{IP3} is higher than the 1-dB compression voltage v_{1dB}; thus, v_L is approximately equal to v_{1dB}. Note that these voltages are all peak differential.

Thus, the linearity of a CMOS differential pair can be adjusted with transistor size and bias current. In comparison, the bipolar differential pair has a fixed v_{IP3} of $4v_T$ or about 100 mV, although it can be improved with resistive degeneration. Note also that the linearity of a differential pair is not equivalent to twice the linearity of a single-ended amplifier since the shape of the transfer curve is quite different and hence the power series expansion is quite different.

7.7 Low Voltage Topologies for LNAs and the Use of On-Chip Transformers

Of the configurations described so far, the common-emitter amplifier would seem ideally suited to low voltage operation. However, if the improved properties of the cascode are required at lower voltage, then the topology must be modified slightly. This has led some designers to "fold" the cascode as shown in Figure 7.37(a) [6]. With the use of two additional LC tanks and one very large coupling capacitor, the cascode can now be operated down to a very low voltage. This approach does have drawbacks, however, as it uses two additional inductors, which will use a lot of die area. The other drawback present with any folding scheme is that both transistors can no longer reuse the current. Thus, this technique will use twice the current of an unfolded cascode, although it could be used at half the voltage, to result in a comparable power consumption.

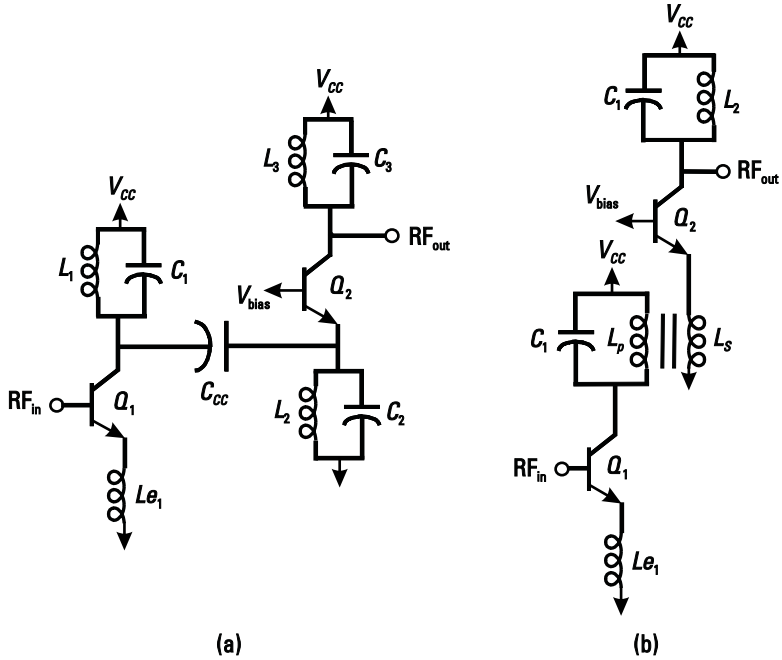

Figure 7.37 A folded cascode LNA with: (a) capacitive coupling and (b) inductive coupling.

An alternative to this topology involves using a transformer to produce magnetic, rather than electric, coupling between the two stages, as shown in Figure 7.37(b). In this circuit, L_p and L_s form the primary and secondary windings of an on-chip transformer, respectively. Note that there is no longer any need for the coupling capacitor. The transformer, although slightly larger than a regular inductor, will nevertheless use much less die area than two individual inductors.

Typically, LNAs, as already discussed, make use of inductors for many reasons including: low loss biasing, maximized signal swing for high dynamic range, and simultaneous noise and power matching. It is also possible to replace the collector and emitter inductors with a transformer as shown in Figure 7.38 [7]. This circuit has all the same useful properties as the previously discussed LNA, but adds some additional benefits. From [7], the gain of this circuit is given by:

$$S_{21} = \frac{g_m Z_L}{A_{\text{BJT}} + g_m Z_L \left(\frac{1}{n} + j\omega r_b C_\mu\right)\left(\frac{1}{n} + 1\right) - \omega^2 L_i C_\mu \left(\frac{1}{n} + 1\right)} \tag{7.136}$$

where

$$A_{\text{BJT}} = 1 + j\omega r_b (C_\pi + C_\mu) - \omega^2 L_i (C_\pi + C_\mu) \tag{7.137}$$

At low frequencies the gain is given by

$$S_{21} \approx \frac{g_m Z_L}{1 + g_m Z_L \left(\frac{1}{n}\right)} \tag{7.138}$$

Under many circumstances, $g_m Z_L$ is large, and the gain is approximately equal to n, the turns ratio. This means that there is very little dependence on transistor parameters.

Considering the redrawn circuit in Figure 7.39, a simplistic description of this circuit can be provided. The transistor acts as a source follower to the input of the transformer. A transformer by itself cannot provide power gain, since, if the voltage

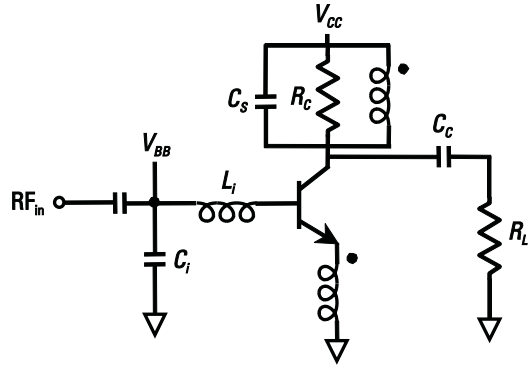

Figure 7.38 LNA with transformer coupling collector to emitter.

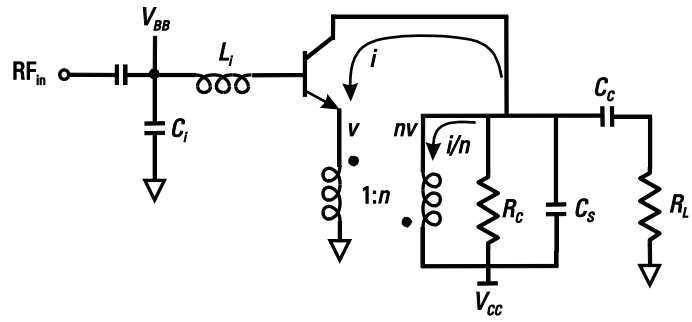

Figure 7.39 Redrawn transformer-coupled LNA.

is increased by a factor of n, the current is decreased by a factor of n. However, in this circuit, the transistor feeds the primary current into the secondary adding it to the secondary current, but also allowing a lower impedance to be driven. The net result is that the gain S_{21} is approximately equal to n. Thus with a turns ratio of 4:1, the amplifier can achieve a gain of 12 dB.

The advantage of this circuit is that the gain is determined largely by the transformer turns ratio, thus minimizing the dependence on transistor parameters. The transformer has high linearity and low noise; thus, the amplifier also has high linearity and low noise. Recently, this type of amplifier and variations of it have also been realized in CMOS with similar benefits of enhanced linearity and low noise with low power supply voltages [8, 9].

7.8 DC Bias Networks

A number of circuits have already been discussed in this text, and it is probably appropriate to say at least a few words about biasing at this point. Bias networks are used in all types of circuits and are not unique to LNAs.

The most common form of biasing in RF circuits is the current mirror. This basic stage is used everywhere and it acts like a current source. Normally, it takes a current as an input and this current is usually generated, along with all other references on the chip, by a circuit called a *bandgap reference generator*. A bandgap reference generator is a temperature independent bias-generating circuit. The bandgap reference generator balances the V_{BE} dependence on temperature, with the temperature dependence of v_T to result in a voltage or current nearly independent of temperature. Design details for the bandgap reference generator will be discussed further in Section 7.8.2 [10].

Perhaps the most basic current mirror is shown in Figure 7.40(a). In this mirror, the bandgap reference generator produces current I_{bias} and forces this current through Q_1. Scaling the second transistor allows the current to be multiplied up and used to bias working transistors. One major drawback to this circuit is that it can inject a lot of noise at the output due primarily to the high g_m of the transistor $N{\cdot}Q_1$ (larger than Q_1 by a factor of N), which acts like an amplifier for noise. A capacitor can be used to clean up the noise and degeneration can be put into the

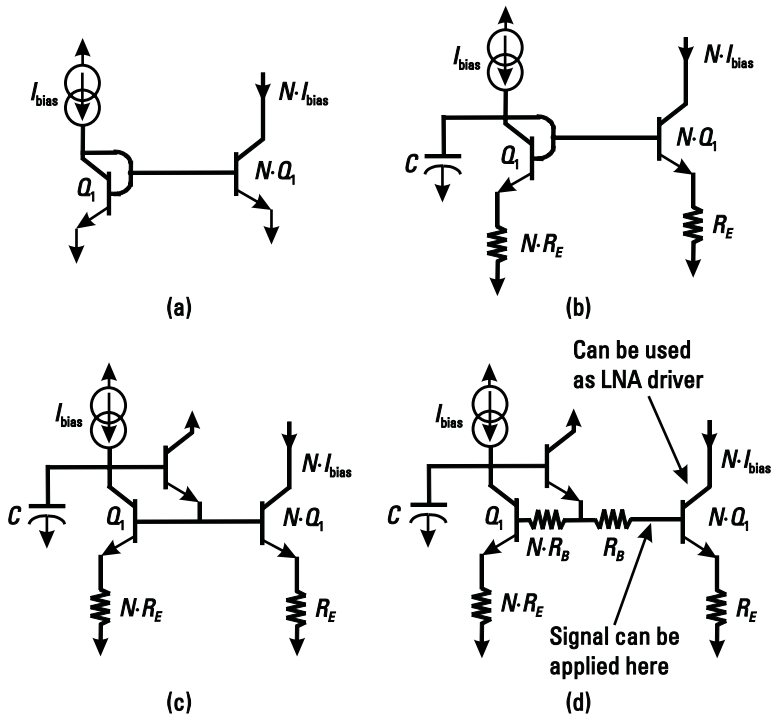

Figure 7.40 Various current mirrors. (a) Simple mirror. (b) Mirror with improved noise performance. (c) Mirror with improved current matching. (d) Mirror with transistor doing double duty as a current source and a driver.

circuit to reduce the gain of the transistor, as shown in Figure 7.40(b). If Q_1 is going to drive many current stages, then base current can affect the matching, so an additional transistor can be added to provide the base current without affecting I_{bias}, as shown in Figure 7.40(c).

Another useful technique for an LNA design is to make the $N{\cdot}Q_1$ transistor function both as a mirror transistor and as the LNA driver transistor, as shown in Figure 7.40(d). In this case, resistors have to be added in the base to isolate the input from the low impedance of Q_1. Provided that R_B is big compared to the input impedance of the transistor $N{\cdot}Q_1$, little noise is injected here.

With any of these mirrors, a voltage at the collector of $N{\cdot}Q_1$ must be maintained above a minimum level, or else the transistor will go into saturation. Saturation will lead to bad matching and to nonlinearity.

7.8.1 Temperature Effects

For transistor current given by:

$$i_C \quad I_S e^{\frac{v_{BE}}{v_T}} \tag{7.139}$$

the temperature affects parameters such as I_S and v_{BE}. As well, the current gain β is affected by temperature. I_S doubles every 10°C rise while the relationship for v_{BE} and β with temperature is shown in (7.140) and (7.141).

$$\left.\frac{v_{BE}}{T}\right|_{i_c=\text{constant}} \quad 2\ \text{mV/°C} \tag{7.140}$$

$$\frac{\beta}{T} \quad +0.5\%/°\text{C} \tag{7.141}$$

A typical temperature range might be 0° to 85°C. Thus for a constant voltage bias, if the current is 1 mA at 20°C, then (7.139) predicts it will change to about 0.2 mA at 0°C and about 71 mA at 85°C, taking into account only the temperature dependence of v_{BE} and v_T. Thus, the current changes by more than 300 times over this temperature range. This illustrates why constant current biasing (for example, with the current mirrors, discussed in Section 7.8) is used. If both transistors in the current mirror are at the same temperature, then output current is roughly independent of temperature.

7.8.2 Temperature Independent Reference Generators

All RF circuits require bias currents, and these currents must be generated somehow. Bias circuits should produce a current or voltage that as much as possible is insensitive to supply and process variations. As well, these references should have some well defined behavior over temperature. This behavior can often be adjusted to help compensate for circuit variation with temperature [1, 11, 12]. Since circuit performance in a silicon process is affected by temperature, supply voltage, and process variations, it is sometimes possible to find dependencies that cancel one another. Most of these types of circuits make use of bipolar transistors, even in an all CMOS process. Typically all CMOS processes include a very slow lateral PNP transistor for the purpose of building bias references.

To start the design, consider the bipolar base emitter voltage characteristic described by

$$V_{BE} = \frac{kT}{q}\ln\frac{I_C}{I_S} \tag{7.142}$$

This expression seems to show that base-emitter voltage is directly proportional to temperature; however, I_S has a large temperature dependence as well. An expression for V_{BE} as a function of temperature is [13, 14]

$$V_{BE} = V_{BG}\ \ 1\ \ \frac{T}{T_O}\div + V_{BEO}\frac{T}{T_O} + \frac{2.3\times kT}{q}\ln\frac{T_O}{T} + \frac{kT}{q}\ln\ \frac{I_C}{I_{CO}}\div \tag{7.143}$$

where T_O is the reference temperature, V_{BEO} is the base-emitter voltage at the reference temperature, and V_{BG} is the bandgap voltage of silicon (approximately 1.206V). Even though it may not be immediately obvious from this expression, V_{BE} will actually decrease for a constant collector current with increasing temperature. Thus, if the collector current is assumed to be constant, then the derivative of V_{BE} with respect to temperature is:

$$\frac{dV_{BE}}{dT} = \frac{V_{BEO} \quad V_{BG}}{T_O} + \frac{2.3\cancel{\times}k}{q} \ln \frac{T_O}{T} + \frac{2.3\cancel{\times}k}{q} + \frac{k}{q} \ln \quad \frac{I_C}{I_{CO}} \div \tag{7.144}$$

Note that this expression shows that not a lot can be done to adjust the slope of the temperature dependence.

Next, to cancel this negative temperature/voltage relationship, a voltage that increases with temperature must be found. To do this, assume that two BJTs are biased at different current densities (one transistor is N times larger than the other), but the same current. Then the difference between their base-emitter voltages will be given by

$$V = V_{BE1} \quad V_{BE2} = \frac{kT}{q} \ln \frac{I_C}{I_{S1}} \quad \frac{kT}{q} \ln \frac{I_C}{I_{S2}} = \frac{kT}{q} \ln \frac{I_{S2}}{I_{S1}} = \frac{kT}{q} \ln N \tag{7.145}$$

Since (7.145) does not have I_S in it, this relationship is much simpler than (7.143) and this voltage is directly proportional to temperature. In this case, the slope of the temperature dependence is given by:

$$\frac{d \quad V}{dT} = \frac{k}{q} \ln \frac{I_{C1}}{I_{C2}} = \frac{k}{q} \ln N \tag{7.146}$$

Therefore, the slope can be adjusted by changing N, the ratio of size of Q_1 and Q_2, and, by a proper choice, can be made equal in magnitude and opposite in sign to (7.145).

A simple circuit that could be used to generate a bandgap reference is shown in Figure 7.41. In this circuit, the opamp is used to keep the voltage at the collector of Q_1 and the voltage at the top of the resistor R_1 the same by adjusting the value of the V_{GS} of the identical PMOS transistors. Thus, the voltage V_{ref} is given by:

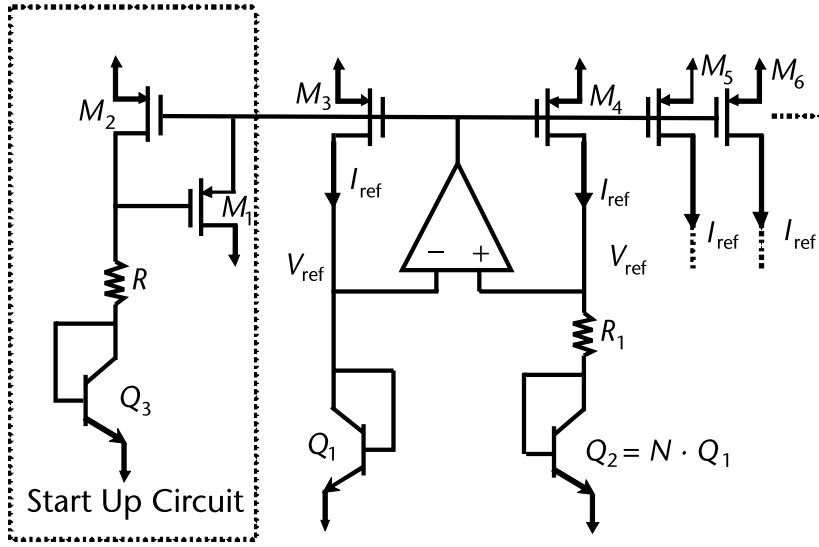

Figure 7.41 A simple bandgap reference generator with a PTAT output current.

$$V_{\text{ref}} = V_{BE1} = V_{BE2} + V_{R1} \tag{7.147}$$

where V_{R1} is the voltage drop across the resistor R_1.

Therefore, the voltage across the resistor R_1 can also be given by:

$$V_{R1} = V_{BE1} \quad V_{BE2} \tag{7.148}$$

Since this voltage is proportional to the difference of two base emitter voltages, V_{ref} is the sum of two voltages that have opposite temperature dependences and consequently this voltage can be made to be independent of temperature. The current through the circuit is given by:

$$I_{\text{ref}} = \frac{V_{BE1} \quad V_{BE2}}{R_1} \tag{7.149}$$

Thus, the current in this circuit I_{ref} is proportional to absolute temperature (PTAT). This current can be scaled or mirrored to other PMOS transistors to create as many copies as is needed.

Note that in practice this circuit requires a start-up circuit. This is because, in addition to the solution just assumed, the condition where all currents in the circuit are zero will also provide a stable operating point for the opamp. One possible start-up circuit is shown in Figure 7.41. When the circuit is powered up, initially there is no current flow. This means that the gate of M_1 will be pulled to ground, which will cause this transistor to act like a short circuit pulling the gate voltage of the rest of the PMOS transistors to ground ensuring that current will start to flow in the circuit. Once the current starts to flow through the circuit, this causes the voltage across the diode transistor Q_3 and the resistor R to rise to the point where M_1 will turn off.

While a PTAT current source is useful for a lot of circuits, another common reference current that is desired is a zero temperature coefficient (ZTC) reference current. Such a reference current can be generated from a modified reference generator as shown in Figure 7.42. In this case the current I_{ref} is the sum of the current flowing

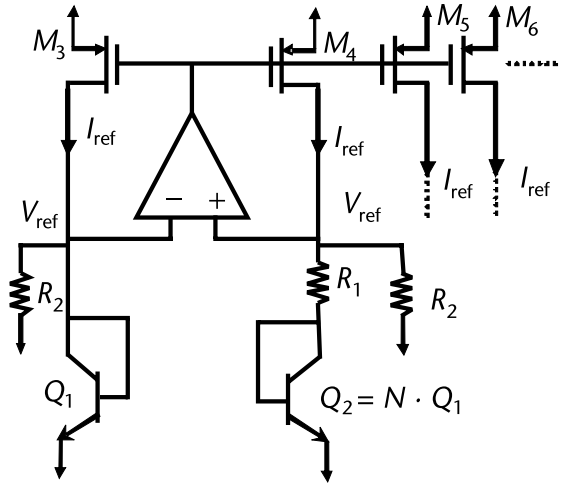

Figure 7.42 A ZTC current generating reference circuit (start-up circuit not shown).

through resistor R_1, which is a PTAT current and the current flowing through R_2. The voltage across R_2 is equal to V_{BE1}; thus, the current I_{ref} is given by:

$$I_{\text{ref}} = \frac{V_{BE1} \quad V_{BE2}}{R_1} + \frac{V_{BE1}}{R_2} \tag{7.150}$$

Since I_{ref} is made up of one component that is proportional to temperature and one that is inversely proportional to temperature, by adjusting the size of these two resistors, the temerature dependence of these two current components can be made to cancel out and sum to a value that is constant over temperature. Using this circuit, it is possible to make a current that will be almost perfectly flat over termperature.

7.8.3 Constant G_M Biasing for CMOS

In CMOS processes sometimes it is desirable to have a constant g_m bias generator. These circuits do not use pn junctions, but rather actual CMOS devices to develop a voltage that is proportional to the difference between two gate source voltages. A circuit that could be used to generate a constant g_m bias is shown in Figure 7.43. In this circuit the current generated is given by:

$$I_{\text{ref}} = \frac{V_{gs1} \quad V_{gs1}}{R_1} \tag{7.151}$$

A common choice for N in this case is 4. The NMOS transistors should have similar parameters to the active RF devices that they are being used to bias. Often this will mean that the device is a short channel device and will suffer from extreme process variation which can be problematic. Also note that in practice, a constant g_m bias will not necessarily mean that a practical amplifier will have a constant gain

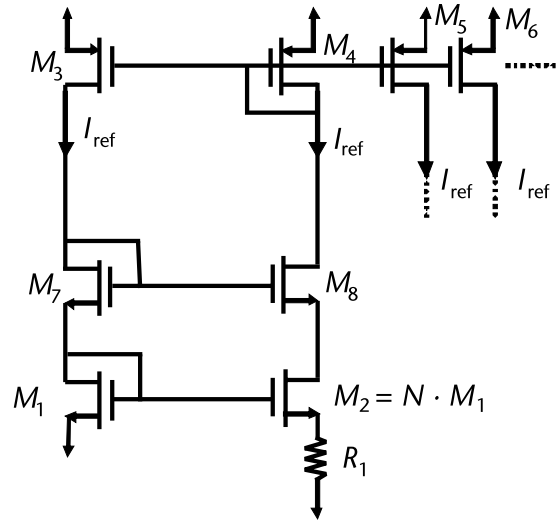

Figure 7.43 A constant g_m current generating reference circuit (start-up circuit not shown).

over temperature. An even more extreme slope to the current temperature curve may be required in practice. In this case it is possible to make a reference current that is the difference of either a constant g_m current and a ZTC current or the difference between a PTAT and a ZTC current. The resulting current will have a much higher voltage/temperature slope than any of the basic references presented here.

7.9 Broadband LNA Design Example

As a major design example, we will design a broadband LNA to work from 50 MHz to 900 MHz, with input matched to 75 and with an S_{11} better than –10 dB over this range. The gain must be more than 12 dB, the noise figure must be less than 5 dB, and the IIP3 must be greater than 6 dBm. The circuit must operate with a 3.3-V supply and consume no more than 8 mA of current. Assume that there is a suitable 50-GHz process available for this design. Also assume that the LNA will drive an on-chip mixer with an input impedance of 5 k .

This is going to be a high linearity part and it needs to be broadband. Therefore, the matching and the load cannot make use of reactive components. This means that, of the designs presented so far, an LNA with shunt feedback will have to be used. This can be combined with a common-base amplifier to provide better frequency response, and an output buffer to avoid the problem of loading the circuit. A first cut at a topology that could satisfy the requirements is shown in Figure 7.44. Note that emitter degeneration has been added to this circuit. Degeneration will almost certainly be required due to the linearity requirement. We have left the current source as ideal for this example. As well, we are not including the bias circuitry that will be needed at the base of Q_1.

The first specification we will satisfy is the requirement of an 8-mA total supply current. It may even be possible to do this design with less current. The trade-off is that as the current is decreased, R_E must be increased to deliver the same linearity, and this will affect noise. We will begin this design using all the allowed current and, at the end, we will consider the possibility of reducing the current in a second iteration.

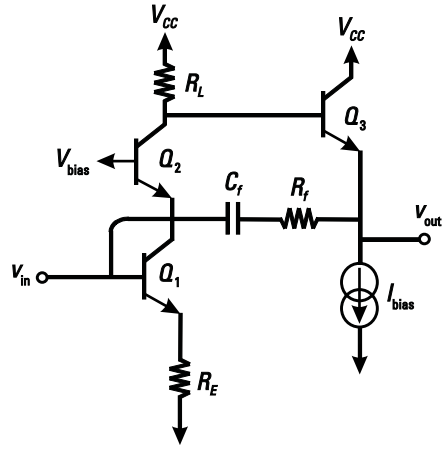

Figure 7.44 Circuit for broadband LNA example.

The total current must be divided between the two stages of this amplifier. The buffer must have enough current so that it continues to operate properly even when it has to deliver a lot of current in the presence of large signals. Since the load resistance is large, the buffer will have to drive an effective resistance of approximately $R_f + 75$. This is expected to be a few hundred ohms, and this will require several hundreds of microamps of ac current. We will start with 3 mA in the buffer and 5 mA in the driver stage.

We can now start to size the resistors and capacitor in the circuit by considering linearity. An IIP3 of 6 dBm in a 75 system, assuming that the input is matched, means that the IIP3 in terms of voltage will be 546 mVrms. Since we have assumed a current of 5 mA in the driver, we can now use (7.95) to determine the size of R_E:

$$R_E = r_e \left(\frac{v_{\text{IP3}}}{2v_T}\right)^{\frac{2}{3}} \quad r_e = 5 \quad \left(\frac{546\ \text{mV}}{2.25\ \text{mV}}\right)^{\frac{2}{3}} \quad 5 \quad = 19.6$$

This is a very rough estimate for what the linearity should be. As well, there are many other factors that can limit the linearity of the circuit. We will start with an $R_E = 20$.

The gain can also be found now. Knowing that we want 12 dB of voltage gain means a gain of 4 V/V. We will assume that the buffer has a voltage gain A_{BO} of about 0.9 (they will always have a bit of loss). Thus, the load resistance can be obtained:

$$G = \frac{R_L}{R_E + r_e} A_{\text{BO}} \qquad R_L = \frac{G}{A_{\text{BO}}}\left(R_E + r_e\right) = \frac{4}{0.9}(20 \quad + 5 \quad) \quad 115$$

Now we need to set the feedback resistor. Knowing that the input impedance needs to be 75 (we approximate that the input impedance is R_f divided by the gain):

$$Z_{\text{in}} \quad \frac{R_f}{\dfrac{R_L}{R_E + r_e}} \qquad R_f = \frac{Z_{\text{in}} R_L}{R_E + r_e} = \frac{75 \quad \times 115}{20 \quad + 5} = 345$$

The other thing that must be set is the value of C_f. Since the LNA must operate down to 50 MHz, this capacitor will have to be fairly large. At 50 MHz if it has an impedance that is one-twentieth of R_f, then this would make it approximately 50 pF. We will start with this value. It can be refined in simulation later.

The only thing left to do in this example is to size the transistors. With all the feedback around this design, the transistors will have a much smaller bearing on the noise figure than in a tuned LNA. Thus, we will make the input transistor fairly large (60 μm) and the other two transistors will be sized to be 30 μm fairly arbitrarily. Having high f_T is important, but in a 50-GHz process, this will probably not be an issue. The other last detail that needs to be addressed is the bias level at the base of Q_2. Given that the emitter of Q_1 is at 100 mV, the base will have to sit at about 1V. The collector of Q_1 should be higher than this, for example, about 1.2V. This means that the base of Q_2 will need to be at about 2.2V and since its collector will sit at about 2.7V, this transistor will have plenty of headroom.

The noise figure of this design can now be estimated. First, the noise voltage produced by the source resistance is:

$$v_{\text{ns}} = \sqrt{4kTR_s} = \sqrt{4\times(4\times10^{-21})\times75} = 1.1\ \text{nV}/\sqrt{\text{Hz}}$$

Since the input is matched, this voltage is divided by half to the input of the driver transistor and then sees the full voltage gain of the amplifier. Thus, the noise at the output due to the source resistance is:

$$v_{o(\text{source})} = \frac{1}{2}v_{\text{ns}}G = \frac{1}{2}\times1.1\ \text{nV}/\sqrt{\text{Hz}}\times4 = 2.2\ \text{nV}/\sqrt{\text{Hz}}$$

The current produced by the degeneration resistor is:

$$i_{n_E} = \sqrt{\frac{4kT}{R_E}} = \sqrt{\frac{4\times(4\times10^{-21})}{20}} = 28.3\ \text{pA}/\sqrt{\text{Hz}}$$

This current is split between the degeneration resistor and the emitter of the driver transistor. The fraction that enters the driver transistor develops into a voltage at the collector of the cascode transistor and is then passed to the output through the follower:

$$v_{\text{on}_E} = i_{n_E}\times\frac{R_E}{r_e+R_E}\div R_L\times A_{\text{BO}} = 28.3\ \text{pA}/\sqrt{\text{Hz}}\times\frac{20}{5+20}\div\times115\times0.9$$

$$= 2.3\ \text{nV}/\sqrt{\text{Hz}}$$

If we assume that the source resistance and the emitter degeneration resistor are the two dominant noise sources, then the noise figure is:

$$\text{NF} = 10\log\frac{v_{\text{on}_E}^{\ 2}+v_{\text{ons}}^{\ 2}}{v_{\text{ons}}^{\ 2}}\div = 10\log\frac{(2.3\ \text{nV}/\sqrt{\text{Hz}})^2+(2.2\ \text{nV}/\sqrt{\text{Hz}})^2}{(2.2\ \text{nV}/\sqrt{\text{Hz}})^2}\div = 3.2\ \text{dB}$$

This performance at the end of this design is verified by simulation. The voltage gain is shown in Figure 7.45 and is between 12.3 and 12.4 dB over the frequency range of interest. This is very close to the value predicted by our calculations and is

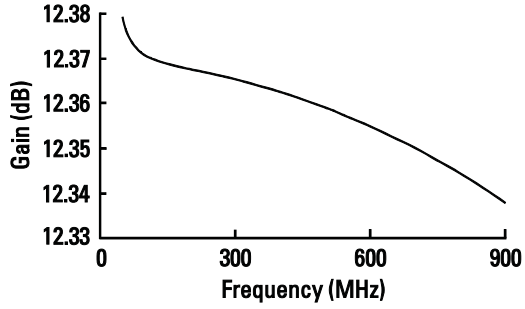

Figure 7.45 Simulated voltage gain of the broadband LNA example circuit.

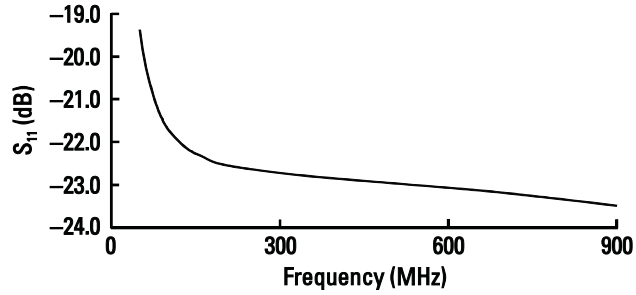

Figure 7.46 Simulated S_{11} of the broadband LNA example circuit.

very constant. The magnitude of S_{11} is shown in Figure 7.46 and is less than –19 dB over the whole range. Thus, the circuit is almost perfectly matched to 75 over all frequencies. The noise figure was also simulated and is shown in Figure 7.47. The noise figure was less than 3.5 dB and only slightly higher than our calculated value. We could have gotten closer to the right value by considering more noise sources. Since this is lower than required, a second iteration of this example could reduce the current in the driver stage and a larger value for R_E could be used to maintain the linearity. In order to test the linearity of the circuit, two tones were fed into the circuit. One was at a frequency of 400 MHz and one at a frequency of 420 MHz, each having an input power of –20 dBm. The *fast Fourier transform* (FFT) of the output is shown in Figure 7.48. The two tones at the fundamental have an rms amplitude of –19.5 dBV and the amplitude of the intermodulation tones have an amplitude of –73.4 dBV. Using (2.55) in Chapter 2, this means that at the input this corresponds to an IIP3 of 572 mV or 6.4 dBm. Thus, the specification for linearity is met for this part.

7.10 Distributed Amplifiers

In this section, distributed amplifiers will be described briefly. More detailed information is available in the literature and in numerous textbooks, including [15, 16]. We will start by describing how distributed amplifiers are based on the artificial transmission line, followed by a description of such transmission lines. The section will end with a summary of the steps in designing distributed amplifiers.

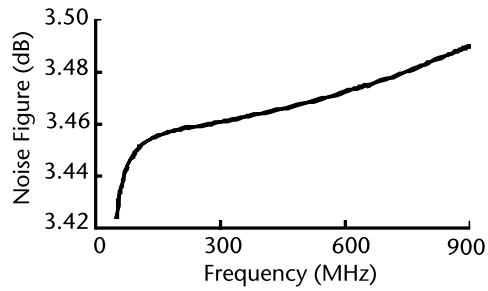

Figure 7.47 Simulated noise figure of the broadband LNA example circuit.

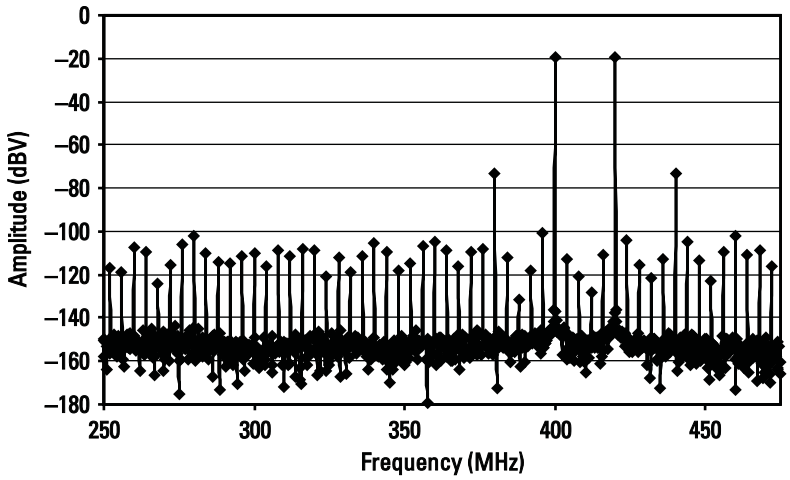

Figure 7.48 FFT of the broadband LNA with two tones applied at the input.

Transmission lines can have flat response to very high frequency in spite of parasitic components, if the line is terminated in its characteristic impedance. Such parasitic components consist of inductance in series and capacitance in parallel. If the line is modeled with a discrete set of inductors and capacitors, such as shown in Figure 7.49, this is referred to as an artificial transmission line.

The distributed amplifier is based on the artificial transmission line where the input and output capacitance of the transistor are absorbed into the parallel capacitance of the line as shown in Figure 7.50 and the connections between the amplifier stages form the series inductance.

The input and output transmission lines are designed such that delays are matched along the two lines. Thus, signals will add in phase. It is possible to get gain to very high frequencies; unity gain bandwidths of over half of the transistor $f_{\max}$ have been reported [17]. The elements shown as inductors can be real inductors or can be made out of lines that behave like inductors. One of the ways to match the delays on the two lines is to use the same unit values for inductance and capacitance. However, typically C_D is smaller than C_G, so additional capacitance C_{add} may be added on the drain line. It is also common to use more complex amplifier structures, for example, the cascode stage or f_T doublers [17]. An important consideration of the distributed amplifier is that to achieve the high frequency response of a transmission line, both the gate line and the drain line must be terminated on the opposite end of the signal source and the load respectively and such terminations are called dummy terminations. The need for the termination, typically 50 , results

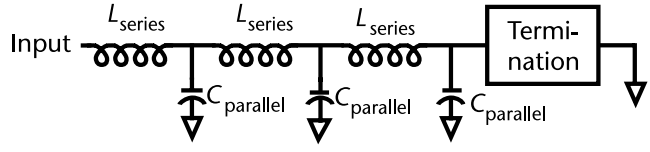

Figure 7.49 Model of a transmission line with termination.

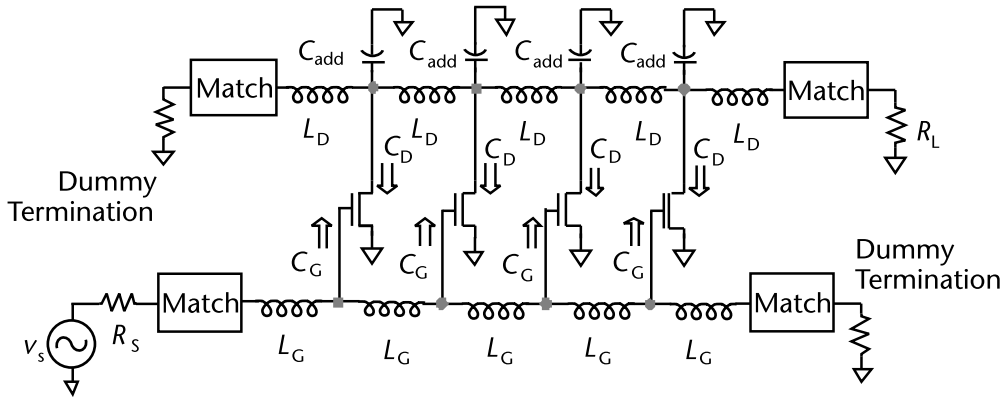

Figure 7.50 A distributed amplifier schematic with input artificial transmission line formed by inductors L_G and capacitors C_G and output artificial transmission lines formed by inductors L_D and capacitors C_D and C_{add}.

in additional power dissipation and noise. If the lines are ac coupled to allow the amplifiers to be locally biased, the lower frequency cutoff is determined by the size of the coupling capacitors.

A brief design procedure for the distributed amplifier will be given in Section 7.10.2, but first, transmission lines will be discussed in the following section as a necessary component of the distributed amplifier.

7.10.1 Transmission Lines

Transmission lines can be modeled as pi, k, or m-derived sections as shown in Figure 7.51.

The characteristic impedance of the transmission line is given by

$$Z_0 = \sqrt{\frac{L}{C}} \tag{7.152}$$

We note that L and C are the per-unit inductance and capacitance. It can be seen that although the values of L and C depend on the number of sections of a

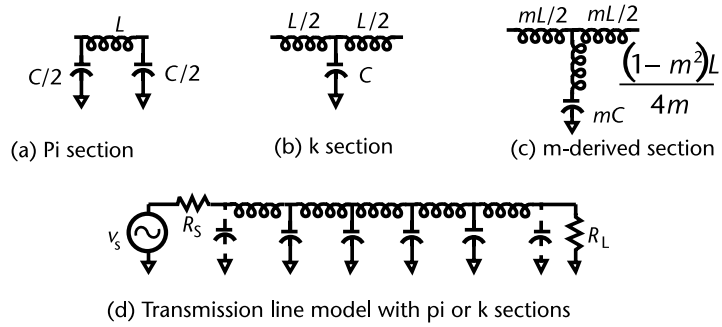

Figure 7.51 Transmission line components: (a) pi section, (b) k section, (c) m-derived section, and (d) model of transmission line with a pi or a k section.

particular line, the ratio remains the same. That is, for N times as many sections, both L and C are reduced to $1/N$ and the ratio remains the same.

The cutoff frequency of the transmission line is given by

$$f_c = \frac{1}{\pi}\sqrt{\frac{1}{LC}} \tag{7.153}$$

Thus, by breaking the line into more sections, as both L and C become smaller, the cutoff frequency is increased. The gain is down by 3 dB at the cutoff frequency, but phase and matching deviate significantly from the ideal transmission line at about 45% of f_c for the pi and k sections; hence, this sets the effective useful frequency range for a particular artificial transmission line. We note that the somewhat more complicated m-derived section is useful up to about 85% of f_c.

If lines are used to form the equivalent series inductance, the following formulas can be used to determine the effective value of L and C, where c is the speed of electromagnetic propagation in free space, ε_{eff} is the effective dielectric constant, Z_0 is the characteristic impedance of the line, and l is the length of the line.

$$L = \frac{Z_0 l\sqrt{\varepsilon_{eff}}}{C} \quad C = \frac{l\sqrt{\varepsilon_{eff}}}{Z_0 c} \tag{7.154}$$

Example 7.13: Modeling a 1-mm, 50 Transmission Line

A 1-mm transmission line of approximately 50 is made with top metal that is 4 μm wide and with an oxide thickness of 4 μm. For this line, derive a model that uses one, two, and five stages. Also, verify the frequency response and matching for this line.

Solution:

For one stage using (7.154), assuming that Z_0 is 50 ,

$$L = \frac{Z_0 l\sqrt{\varepsilon_{\text{eff}}}}{C} = \frac{50 \quad 1 \quad 10^{-3}\sqrt{3.9}}{3 \quad 10^8} = 329 \text{ pH}$$

$$C = \frac{l\sqrt{\varepsilon_{\text{eff}}}}{Z_o C} = \frac{1 \quad 10^{-3}\sqrt{3.9}}{50 \quad 3 \quad 10^8} = 132 \text{ fF}$$

Check:

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{329 \text{ pH}}{132 \text{ fF}}} = 49.92$$

The cutoff frequency can be determined from (7.153),

$$\sqrt{\frac{1}{LC}} = \frac{1}{\pi}\sqrt{\frac{1}{329 \text{ pH} \quad 132 \text{ fF}}} = 48.3 \text{ GHz}$$

For two or five stages, L and C are simply divided by 2 and by 5, respectively, for which the resulting cutoff frequency will go up by a factor of 2 and 5 to 96.6 GHz and 241.5 GHz, respectively. The resulting schematics were simulated, with results shown in Figure 7.52.

It can be seen that as predicted, the 3-dB frequency occurs at roughly the calculated f_c; however, the phase response, at least for the higher-order models, is deviating from the ideal well before the cutoff frequency. Matching shown in Figure 7.52(c) can be seen to be good, with S_{11} lower than –15 dB, up to about 45% of f_c. Thus, with one section, the model is good up to about 21.5 GHz. With two stages, matching is good to about 43 GHz, while with five stages, matching is good to about 109 GHz.

7.10.2 Steps in Designing the Distributed Amplifier

The following are steps in designing the distributed amplifier:

1. The transmission line characteristic impedance, Z_0, from (7.152) is typically adjusted to match the source and load. In a fully integrated context such impedances can be higher than 50 in order to save power. In some cases, a deliberate mismatch of up to 30% between the drain and the gate line cutoff frequency is introduced to control overshoot. This step sets the ratio of L to C.
2. In this step the absolute values of L and C are determined from the artificial transmission line cutoff f_c given by (7.153) so that the maximum useful

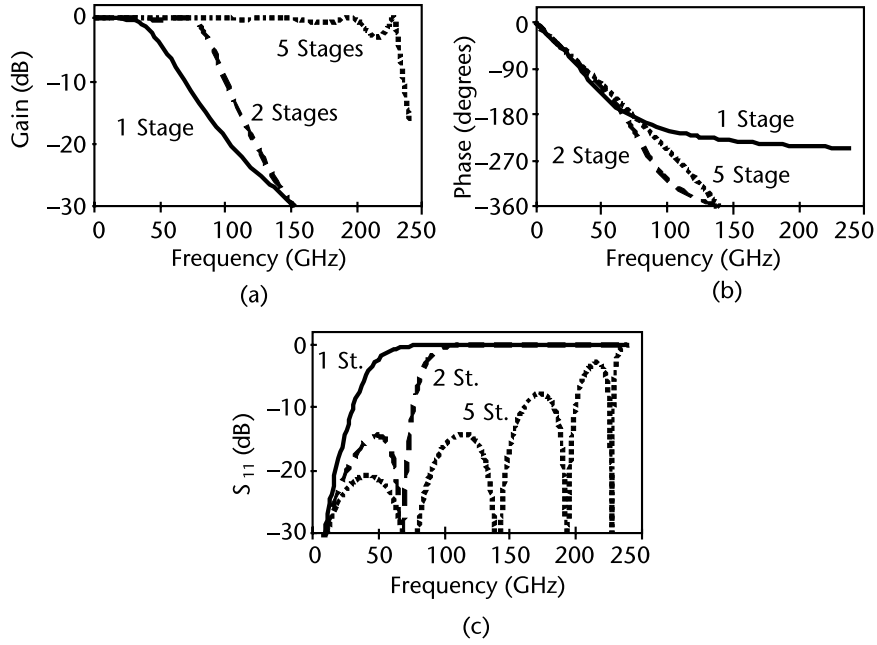

Figure 7.52 Results for one, two, or five stages for a 1-mm transmission line terminated with 50 : (a) gain versus frequency, (b) phase versus frequency, and (c) matching (S_{11}) versus frequency.

frequency is approximately equal to or greater than the maximum operating frequency. Note that the maximum useful frequency is about 45% of f_c for k sections and about 85% of f_c for m-derived sections. Note that f_{max} is the upper limit for the maximum operating frequency.

3. The line capacitance comes from both the transmission line parasitic capacitance and the transistor parasitic capacitance. The transistors are sized based on allowed capacitance and the required g_m to achieve the desired gain. Note that more stages can be used to increase the gain until the losses due to the longer line offset the increase of gain. Due to the fairly high loss in a standard CMOS process, the optimum number of stages is typically about 4 or 5 [16, 17]. However, the use of negative resistance circuits to provide loss compensation has allowed some designs with more stages and potentially with higher gain [18].
4. The transistors are then optimized for highest f_{max} with current level and use of multiple fingers to reduce series gate resistance. In order to optimize f_{max}, the gate finger size should be smaller than some specified maximum size, typically a few microns.
5. Simulations and iterations are performed to verify calculations, to include parasitic capacitances, and to optimize performance.

Note that if the transistor size is made larger, the capacitance will be increased. To keep the line impedance the same, the inductance would also need to be increased, and as a result, the line cutoff frequency, and hence the maximum possible operating frequency, would be reduced. However, with a larger transistor, the optimal current will be higher, resulting in a higher g_m and hence a higher gain. Thus, we can see that it is possible to trade off between bandwidth and gain. If gain, in fact, is the required parameter, W/L is fixed by the need for a particular g_m, and this sets the line capacitance. Hence, inductance can be determined from the desired line impedance.

References

[1] Johns, D. A., and K. Martin, *Analog Integrated Circuit Design*, New York: John Wiley & Sons, 1997.

[2] Voinigescu, S. P., et al., "A Scalable High-Frequency Noise Model for Bipolar Transistors with Applications to Optimal Transistor Sizing for Low-Noise Amplifier Design," *IEEE J. Solid-State Circuits*, Vol. 32, September 1997, pp. 1430–1439.

[3] van der Heijden, M. P., H. C. de Graaf, and L. C. N. de Vreede, "A Novel Frequency Independent Third-Order Intermodulation Distortion Cancellation Technique for BJT Amplifiers," *Proc. BCTM,* September 2001, pp. 163–166.

[4] Toole, B., C. Plett, and M. Cloutier, "RF Circuit Implications of Moderate Inversion Enhanced Linear Region in MOSFETs," *Trans. Circuits and Systems I*, Vol. 51, February 2004, pp. 319–328.

[5] Gonzalez, G., *Microwave Transistor Amplifiers,* 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 1997.

[6] Ray, B., et al., "A Highly Linear Bipolar 1V Folded Cascode 1.9GHz Low Noise Amplifier," *Proc. BCTM*, September 1999, pp. 157–160.

[7] Long, J. R., and M. A. Copeland, "A 1.9GHz Low-Voltage Silicon Bipolar Receiver Front-End for Wireless Personal Communications Systems," *IEEE J. Solid-State Circuits*, Vol. 30, December 1995, pp. 1438–1448.

[8] Cassan, D. J., and J. R. Long, "A 1-V Transformer-Feedback Low-Noise Amplifier for 5-GHz Wireless LAN in 0.18-μm CMOS," *IEEE J. Solid-State Circuits*, Vol. 38, March. 2003, pp. 427–435.

[9] Reiha, M. T., and J. R. Long, "A 1.2 V Reactive-Feedback 3.1–10.6 GHz Low-Noise Amplifier in 0.13 μm CMOS," *IEEE J. Solid-State Circuits*, Vol. 42, May 2007, pp. 1023–1033.

[10] Gray, P. R., et al., *Analysis and Design of Analog Integrated Circuits,* 4th ed., New York: John Wiley & Sons, 2001.

[11] Razavi, B., *Design of Analog CMOS Integrated Circuits*, New York: McGraw-Hill, 2001.

[12] Lee, T. H., *The Design of CMOS Radio Frequency Integrated Circuits*, 2nd ed., Cambridge, U.K.: Cambridge University Press, 2004.

[13] Brugler, J., "Silicon Transistor Biasing for Linear Collector Current Temperature Dependence," *IEEE J. Solid-State Circuits*, Vol. SC-2, June 1967, pp. 57–58.

[14] Tsividis, Y., "Accurate Analysis of Temperature Effects in I_C-V_{BE} Characteristics with Application to Bandgap Reference Sources," *IEEE J. Solid-State Circuits*, Vol. 15, December 1980, pp. 1076–1084.

[15] Pozar, D. M., *Microwave Engineering*, 2nd ed., New York: John Wiley & Sons, 1998.

[16] Wong, T. T. Y., *Fundamentals of Distributed Amplification*, Norwood, MA: Artech House, 1993.

[17] Amaya, R. E., J. Aguirre, and C. Plett, "Gain Bandwidth Considerations in Fully Integrated Distributed Amplifiers Implemented in Silicon," *Proc. IEEE Int. Symp. Circuits and Systems*, Vancouver, Canada, May 2004, Vol. 4, pp. 273–276.

[18] Moez, K., and M. Elmasry, "A 10dB 44GHz Loss-Compensated CMOS Distributed Amplifier," *Proc. Int. Solid-State Circuits Conf.* San Francisco, CA, February 2007, pp. 548–549.

Selected Bibliography

Abou-Allam, E., J. J. Nisbet, and M. C. Maliepaard, "A 1.9GHz Front-End Receiver in 0.5μm CMOS Technology," *IEEE J. Solid-State Circuits*, Vol. 36, October 2001, pp. 1434–1443.

Baumberger, W., "A Single-Chip Rejecting Receiver for the 2.44 GHz Band Using Commercial GaAs-MESFET-Technology," *IEEE J. Solid-State Circuits*, Vol. 29, October 1994, pp. 1244–1249.

Copeland, M. A., et al., "5-GHz SiGe HBT Monolithic Radio Transceiver with Tunable Filtering," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 48, February 2000, pp. 170–181.

Harada, M., et al., "2-GHz RF Front-End Circuits at an Extremely Low Voltage of 0.5V," *IEEE J. Solid-State Circuits*, Vol. 35, December 2000, pp. 2000–2004.

Krauss, H. L., C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, New York: John Wiley & Sons, 1980.

Long, J. R., "A Low-Voltage 5.1-5.8GHz Image-Reject Downconverter RFIC," *IEEE J. Solid-State Circuits*, Vol. 35, September 2000, pp. 1320–1328.

Macedo, J. A., and M. A. Copeland, "A 1.9 GHz Silicon Receiver with Monolithic Image Filtering," *IEEE J. Solid-State Circuits*, Vol. 33, March 1998, pp. 378–386.

Razavi, B., "A 5.2-GHz CMOS Receiver with 62-dB Image Rejection," *IEEE J. Solid-State Circuits*, Vol. 36, May 2001, pp. 810–815.

Rogers, J. W. M., J. A. Macedo, and C. Plett, "A Completely Integrated Receiver Front-End with Monolithic Image Reject Filter and VCO," *Proc. IEEE RFIC Symposium*, June 2000, pp. 143–146.

Rudell, J. C., et al., "A 1.9-GHz Wide-Band IF Double Conversion CMOS Receiver for Cordless Telephone Applications," *IEEE J. Solid-State Circuits*, Vol. 32, December 1997, pp. 2071–2088.

Samavati, H., H. R. Rategh, and T. H. Lee, "A 5-GHz CMOS Wireless LAN Receiver Front End," *IEEE J. Solid-State Circuits*, Vol. 35, May 2000, pp. 765–772.

Schmidt, A., and S. Catala, "A Universal Dual Band LNA Implementation in SiGe Technology for Wireless Applications," *IEEE J. Solid-State Circuits*, Vol. 36, July 2001, pp. 1127–1131.

Schultes, G., P. Kreuzgruber, and A. L. Scholtz, "DECT Transceiver Architectures: Superheterodyne or Direct Conversion?" *Proc. 43rd Vehicular Technology Conference*, Secaucus, NJ, May 18–20, 1993, pp. 953–956.

Steyaert, M., et al., "A 2-V CMOS Transceiver Front-End," *IEEE J. Solid-State Circuits*, Vol. 35, December 2000, pp. 1895–1907.

Yoshimasu, T., et al., "A Low-Current Ku-Band Monolithic Image Rejection Down Converter," *IEEE J. Solid-State Circuits*, Vol. 27, October 1992, pp. 1448–1451.

CHAPTER 8

Mixers

8.1 Introduction

The purpose of the mixer is to convert a signal from one frequency to another. In a receiver, this conversion is from radio frequency to an intermediate frequency, or to baseband for a direct conversion receiver. In a transmitter, this conversion is from baseband or some intermediate frequency up to the radio frequency. Mixing requires a circuit with a nonlinear transfer function, since nonlinearity is fundamentally necessary to generate new frequencies. As described in Chapter 2, if an input RF signal and a local oscillator signal are passed through a system with a second-order nonlinearity, the output signals will have components at the sum and difference frequencies. A circuit realizing such nonlinearity could be as simple as a diode followed by some filtering to remove unwanted components. On the other hand, it could be more complex; such as the double-balanced cross-coupled circuit, commonly called the *Gilbert cell*. In an integrated circuit, the more complex structures are often preferred, since extra transistors can be used with little extra cost but with improved performance. In this chapter, the focus will be on the cross-coupled double-balanced mixer. Consideration will also be given as to how to design a mixer in a low voltage process.

8.2 Mixing with Nonlinearity

A diode or a transistor can be used as a nonlinearity. The two signals to be mixed are combined and applied to the nonlinear circuit. In the transistor, they can be applied separately to two control inputs, for example to the base and emitter in a bipolar transistor or to the gate and source in a field-effect transistor. If a diode is the nonlinear device, then signals might be combined with additional circuitry. As described in Chapter 2, for two inputs at ω_1, and ω_2, with amplitudes A_1, A_2, that are passed through a nonlinearity that multiplies the two signals together will produce mixing terms at $\omega_1 \pm \omega_2$. In addition, other terms (harmonics, feedthrough, intermodulation) will be present and will need to be filtered out.

8.3 Basic Mixer Operation

Mixers can be made from the LNAs that have already been discussed and some form of controlled inverter. One of the simplest forms of a mixer is shown in

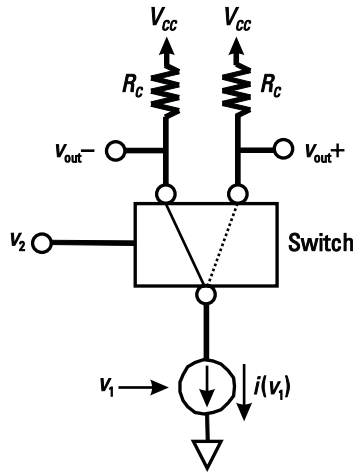

Figure 8.1 Simple conceptual schematic of a mixer.

Figure 8.1. The input of the mixer is simply a gain stage like one that has already been considered. The amplified current from the gain stage is then passed into the switching stage. This stage steers the current to one side of the output or the other depending on the value of v_2 (this provides the nonlinearity just discussed). If the control signal is assumed to be a periodic one, then this will have the effect of multiplying the current coming out of the gain stage by plus or minus one (a square wave). Multiplying a signal by another signal will cause the output to have components at other frequencies. Thus, this can be used to move the signal v_1 from one frequency to another.

8.4 Transconductance-Controlled Mixer

A transconductance-controlled mixer can be made from a bipolar or CMOS differential pair. These were discussed in Chapter 7 when used as differential amplifiers. These are shown along with their basic equations in Figures 8.2 and 8.3. For small signals ($|v_2| < v_L$ shown in Figures 8.2 and 8.3), the current is related to the input voltage v_2 by the transconductance of the input transistors Q_1 and Q_2, or M_1 and M_2. However, the transconductance is controlled by the current I_o, which in turn is controlled by the input voltage v_1. Thus, the output current will be dependent on both input voltages v_1 and v_2.

Let us now look in detail at the operation of the mixers. As shown in Chapter 7, the current in a bipolar differential pair is related to the voltage by the following equation:

$$i_1 = \frac{I_o}{1 + e^{\,v_2 / v_T}}$$
$$i_2 = \frac{I_o}{1 + e^{v_2 / v_T}} \tag{8.1}$$

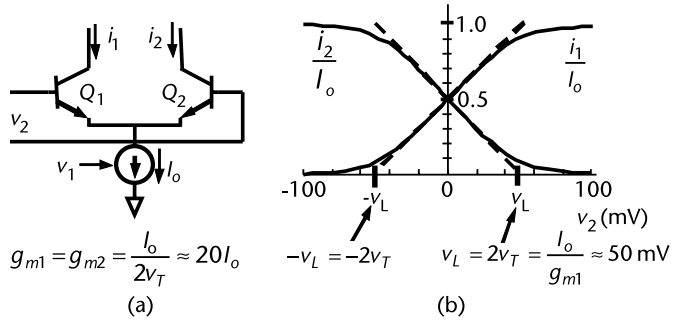

Figure 8.2 Bipolar transconductance controlled mixer: (a) basic circuit and (b) currents.

Thus, the difference in the output currents from the mixer is given by

$$i_o = i_1 \quad i_2 = I_o \quad \frac{1}{1+e^{\ v_2/v_T}} \quad \frac{1}{1+e^{v_2/v_T}} \div = I_o \tanh\frac{v_2}{2v_T} \tag{8.2}$$

This can be converted to a differential voltage with equal load resistors in the collectors.

Now, if we assume the current source is modulated by v_1 so that the current is $I_o + g_{mc}v_1$, where g_{mc} is the transconductance of the current source, then

$$i_o = (I_o + g_{\rm mc}v_1)\tanh\frac{v_2}{2v_T} = \underbrace{I_0\tanh\frac{v_2}{2v_T}}_{v_2 \text{ feedthrough}} + \underbrace{g_{\rm mc}v_1\tanh\frac{v_2}{2v_T}}_{\text{multiplication (mixing)}} \tag{8.3}$$

This equation shows that there will be feedthrough from v_2, whereas an ideal multiplier would respond only to the product of v_1 and v_2. We note that feedthrough from v_1 will appear equally in the output voltages above (common mode) and so does not appear in the differential output voltage.

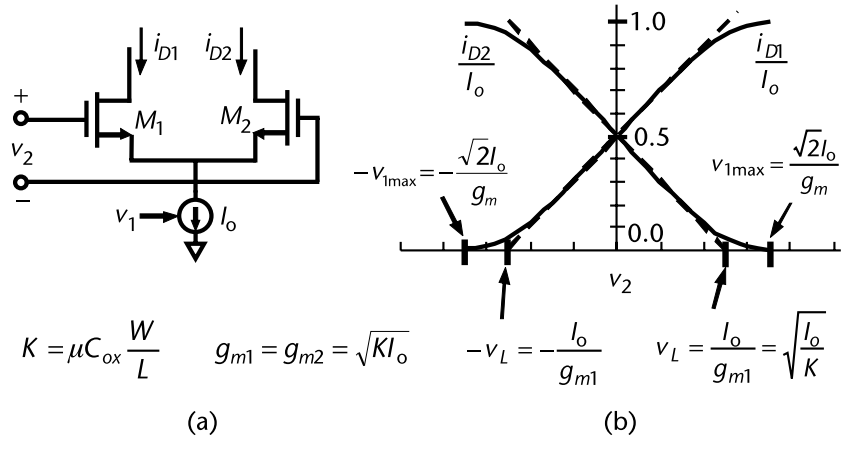

Figure 8.3 CMOS transconductance controlled mixer: (a) basic circuit and (b) currents.

For a CMOS differential pair, from Chapter 7, the currents are:

$$i_{D1,2} = \frac{I_o}{2} \pm \frac{g_m v_2}{2}\sqrt{1 - \frac{(\mu C_{ox})}{4I_o}\frac{W}{L} v_2^2} \tag{8.4}$$

where

$$g_m = \sqrt{\mu C_{ox}\frac{W}{L} I_o} = \sqrt{KI_o} \tag{8.5}$$

where K gathers the process constants and transistor size as

$$K = \mu C_{ox}\frac{W}{L} \tag{8.6}$$

Thus, the difference in the output currents from the mixer is given by

$$i_o = i_{D1} - i_{D2} = g_m v_2\sqrt{1 - \frac{(\mu C_{ox})}{4I_o}\frac{W}{L} v_2^2} = \sqrt{KI_o}\, v_2\sqrt{1 - \frac{(\mu C_{ox})}{4I_o}\frac{W}{L} v_2^2} \tag{8.7}$$

For a large v_2, (8.7) can be simplified noting that the current will flow almost completely through one side of the differential pair or the other:

$$i_o \approx I_o \,\mathrm{sgn}(v_2) \tag{8.8}$$

where $\mathrm{sgn}(v_2)$ will be equal to +1 if v_2 is positive and −1 if v_2 is negative.

Now, if we again assume the current source is modulated by v_1 so that the current is $I_o + g_{mc}v_1$, where g_{mc} is the transconductance of the current source, then

$$i_o = (I_o + g_{mc}v_1)\mathrm{sgn}(v_2) = \underbrace{I_0 \mathrm{sgn}(v_2)}_{v_2 \text{ feedthrough}} + \underbrace{g_{mc}v_1 \mathrm{sgn}(v_2)}_{\text{multiplication (mixing)}} \tag{8.9}$$

8.5 Double-Balanced Mixer

In order to eliminate the v_2 feedthrough, it is possible to combine the output of this circuit with another circuit driven by $-v_2$ as shown with bipolar transistors in Figure 8.4. This circuit has four switching transistors know as the *switching quad*. The output current from the second differential pair is given by:

$$i_o = -i_6 - i_5 = I_o \tanh\frac{v_2}{2v_T} - g_{mc}v_1 \tanh\frac{v_2}{2v_T} \tag{8.10}$$

Therefore, the total differential current is:

$$i_{ob} = i_o - i_o = 2g_{mc}v_1 \tanh\frac{v_2}{2v_T} \tag{8.11}$$

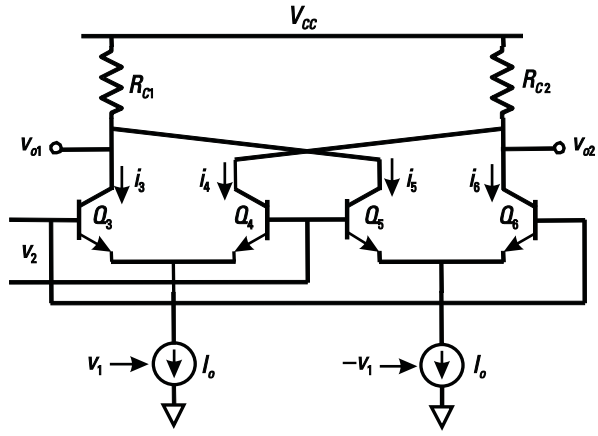

Figure 8.4 Double-balanced transconductance-controlled mixer.

This removes the v_2 feedthrough term that was present in (8.3).

The last step to making this circuit practical is to replace the ideal current sources with an actual amplifier stage, as shown in Figure 8.5. Now v_1 is applied to a differential pair so that the small-signal component of i_1 and i_2 are the inverse of each other; that is,

$$i_1 \quad I_o + \frac{v_1}{2}\frac{1}{r_e + R_E} \tag{8.12}$$

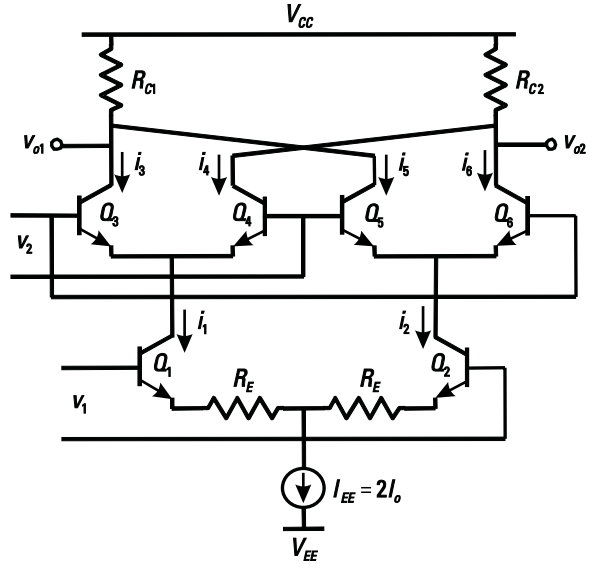

Figure 8.5 Double-balanced mixer with amplifier implemented as a differential pair with degeneration.

and

$$i_2 = I_o \quad \frac{v_1}{2}\frac{1}{r_e + R_E} \tag{8.13}$$

Currents from the switching quad are related to v_2, i_1, and i_2 by (8.2) through (8.11).

$$\begin{aligned} i_3 &= \frac{i_1}{1 + e^{\ v_2 / v_T}} \\ i_4 &= \frac{i_1}{1 + e^{v_2 / v_T}} \\ i_5 &= \frac{i_2}{1 + e^{v_2 / v_T}} \\ i_6 &= \frac{i_2}{1 + e^{\ v_2 / v_T}} \end{aligned} \tag{8.14}$$

Then, assuming that the amplifier formed by Q_1 and Q_2 is linear,

$$(i_3 + i_5) \quad (i_4 + i_6) = \tanh \ \frac{v_2}{2v_T} \div (i_1 \quad i_2) = \tanh \ \frac{v_2}{2v_T} \div \frac{v_1}{r_e + R_E} \tag{8.15}$$

The output differential voltage is

$$v_o = \quad \tanh \ \frac{v_2}{2v_T} \div \frac{v_1}{r_e + R_E} R_C \tag{8.16}$$

Thus, the gain of the circuit relative to v_1 can be determined:

$$\frac{v_o}{v_1} = \quad \tanh \ \frac{v_2}{2v_T} \div \frac{R_C}{r_e + R_E} \tag{8.17}$$

with $R_E = 0$, a general large-signal expression for the output can also be written:

$$v_o = \quad 2R_C I_{\mathrm{EE}} \tanh \ \frac{v_1}{2v_T} \div \tanh \ \frac{v_2}{2v_T} \div \tag{8.18}$$

This takes into account nonlinearity from the bottom pair as well as from the top quad. Previously with R_E present, the bottom was treated in a linear fashion.

For the equivalent CMOS mixer as shown in Figure 8.6, the transconductance of the upper quad is no longer directly proportional to the current (for small signals) but roughly to the square root of current. Hence the CMOS equations are more complex when considered as a multiplier, and the transfer functions will not be given here. This is considered acceptable since, as discussed in the next section, mixers are usually operated with the upper quad being switched, and equations will be given in that section for both bipolar and CMOS mixers.

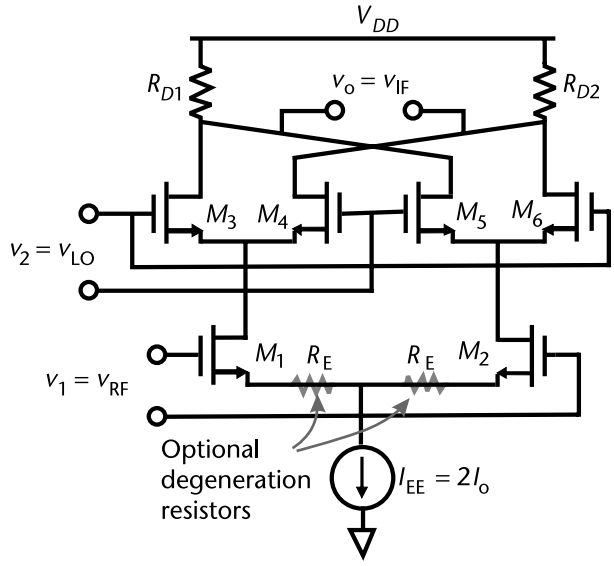

Figure 8.6 CMOS double-balanced mixer.

8.6 Mixer with Switching of Upper Quad

Usually a downconverting mixer is operated with v_1 as the RF signal and v_2 as the local oscillator. The RF input must be linear, and linearity is usually improved by degeneration resistors in the case of bipolar mixers. For CMOS mixers, adjusting transistor sizes and currents can be used to adjust linearity. However, in practice also using degeneration resistors may prove to be more reliable in the presence of process, power supply voltage, and temperature variations. Why must the RF input be linear? In a communications receiver application, the RF input can have several channels at different frequencies and different amplitudes. If the RF input circuitry were nonlinear, adjacent channels could intermodulate and interfere with the desired channel. For example, with inputs at 900, 900.2, 900.4 MHz, if 900 MHz is desired, the third-order intermodulation term from the two other signals occurs at 2 900.2 900.4 = 900 MHz, which is directly on top of the desired signal.

The LO input need not be linear, since the LO is clean and of known amplitude. In fact, the LO input is usually designed to switch the upper quad so that for half the cycle Q_3 and Q_6 (or M_3 and M_4 for the CMOS mixer) are on, and taking all of the current i_1. For the other half of the LO cycle, Q_3 and Q_6 are off and Q_4 and Q_5 (or M_4 and M_5 for the CMOS mixer) are on. More formally, if $v_2 >> 2v_T$, then (8.17) can be approximated as

$$\frac{v_o}{v_1} = u(v_2)\frac{R_C}{r_e + R_E} \tag{8.19}$$

where

$$u(v_2) = \begin{matrix} 1 & \text{if } v_2 \text{ is positive} \\ 1 & \text{if } v_2 \text{ is negative} \end{matrix} \tag{8.20}$$

This is equivalent to alternately multiplying the signal by 1 and 1. For the CMOS mixer, this becomes

$$\frac{v_{IF}}{v_{RF}} = u(v_2) g_{m1} R_D \tag{8.21}$$

If the CMOS mixer had degeneration resistors R_E, the equation would be

$$\frac{v_{IF}}{v_{RF}} = u(v_2) \frac{R_D}{1/g_{m1} + R_E} \tag{8.22}$$

Note that it is assumed that the current to voltage relationship in the RF input is linear. As discussed earlier, this is achieved in a bipolar mixer with appropriate degeneration and in a CMOS mixer with appropriate sizing and biasing of the RF input transistors, and sometimes with degeneration.

This can also be expressed as a Fourier series. If v_2 is a sine wave with frequency ω_{LO} then:

$$u(v_2) = \frac{4}{\pi}\sin(\omega_{\mathrm{LO}} t) + \frac{4}{3\pi}\sin(3\omega_{\mathrm{LO}} t) + \frac{4}{5\pi}\sin(5\omega_{\mathrm{LO}} t) + \frac{4}{7\pi}\sin(7\omega_{\mathrm{LO}} t) + \cdots \tag{8.23}$$

8.6.1 Why LO Switching?

For small LO amplitude, the amplitude of the output depends on the amplitude of the LO signal. Thus, gain is larger for larger LO amplitude. For large LO signals, the upper quad switches and no further increases occur. Thus, at this point, there is no longer any sensitivity to LO amplitude.

As the LO is tuned over a band of frequencies, for example, to pick out a channel in the 902- to 928-MHz range, the LO amplitude may vary. If the amplitude is large enough, the variation does not matter because as shown in Figures 8.2 and 8.3 a large LO will completely switch the current to one side or the other regardless of its amplitude.

For image-reject mixers (to be discussed in Section 8.11), matching two LO signals in amplitude and phase is important. By using a switching modulator, and feeding the LO signal into the switching input, amplitude matching is less important.

Noise is minimized with large LO amplitude. With large LO, the upper quad transistors are alternately switched between completely off and fully on. When off, the transistor contributes no noise and when fully on, the switching transistor behaves as a cascode transistor, which, as described in Chapter 7, does not contribute significantly to noise.

8.6.2 Picking the LO Level

For a bipolar mixer, as shown in Figure 8.2, the differential pair will require an input voltage swing of about 4 5 v_T for the transistors to be hard switched one way or the other. Therefore, the LO input to the mixer should be at least 100-mV peak for complete switching. At 50 , 100-mV peak is 10 dBm. Small improvements in noise figure and conversion gain can be seen for larger signals; however for LO levels

larger than about 0 dBm, there is minimal further improvement. Thus, 10 to 0 dBm (100–300 mV) is a reasonable compromise between noise figure, gain, and required LO power. If the LO voltage is made too large, then a lot of current has to be moved into and out of the bases of the transistors during transitions. This can lead to spikes in the signals and can actually reduce the switching speed and cause an increase in LO feedthrough. Thus, too large a signal can be just as bad as too small a signal.

For a CMOS mixer, as shown in Figure 8.3, required switching levels depends on the ratio of bias current to transconductance. To reduce the required LO amplitude, typically large transistors with low current densities are chosen allowing input amplitudes also in the range of a few hundred millivolts. The trade-off is that as the transistor current density is reduced, the f_T of the transistor is also reduced. When f_T is very important, for example operation at high frequencies, there will be no choice but to increase the LO amplitude in order to get full switching.

Another concern is the parasitic capacitance on node V_d, as shown in Figure 8.7. The transistors have to be turned on and off, which means that any capacitance in the emitter has to be charged and discharged. Essentially, the input transistors behave like a simple rectifier circuit as shown in Figure 8.7(b). If the capacitance on the emitters is too large, then V_d will stop following the input voltage and the transistors will start to be active for a smaller portion of the cycle, as shown in Figure 8.7(c). Since V_d is higher than it should be, it takes longer for the transistor to switch and it switches for a smaller portion of the cycle. This will lead to waveform distortion. This effect, shown for the bipolar case, applies to both bipolar and CMOS.

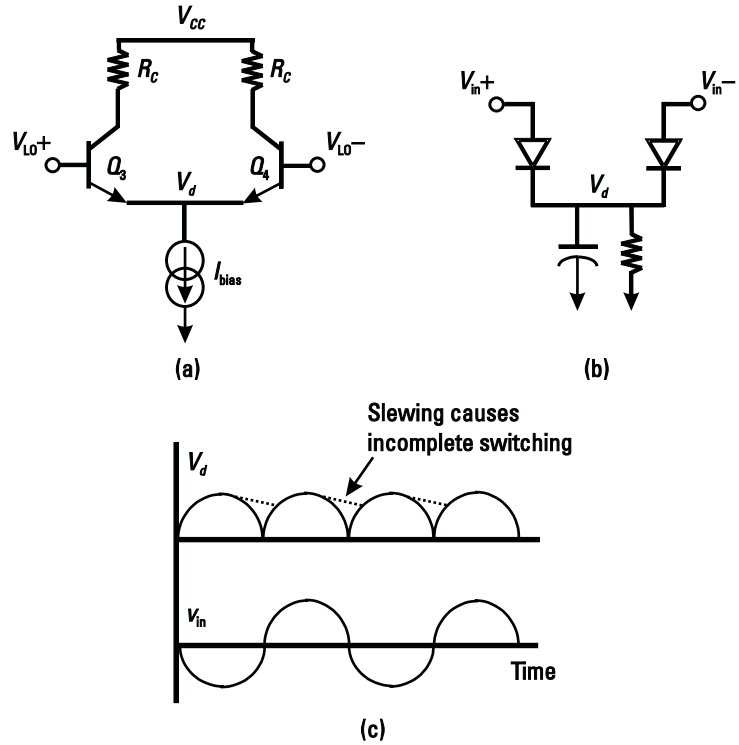

Figure 8.7 Large-signal behavior of the differential pair: (a) schematic representation; (b) diode rectifier model, and (c) waveforms illustrating the problem of slewing.

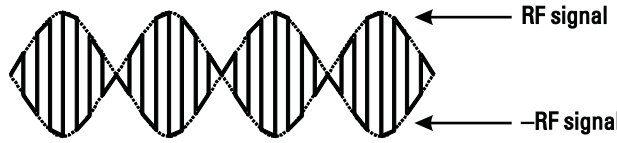

Figure 8.8 Switching waveform.

8.6.3 Analysis of Switching Modulator

The top switching quad alternately switches the polarity of the output signal as shown in Figure 8.8.

The LO signal has the effect of multiplying the RF input by a square wave going from 1 to +1. In the frequency domain, this is equivalent to a convolution of RF and LO signals, which turns out to be a modulation of the RF signal with each of the Fourier components of the square wave.

As can be seen from Figure 8.9, the output amplitude of the product of the fundamental component of square wave is:

$$\begin{aligned} v_o &= \frac{4}{\pi} v_{\mathrm{RF}} \sin(\omega_{\mathrm{RF}} t) \sin(\omega_{\mathrm{LO}} t) \\ &= \frac{1}{2} \times \frac{4}{\pi} v_{\mathrm{RF}} \sin[(\omega_{\mathrm{RF}} + \omega_{\mathrm{LO}}) t] + \frac{1}{2} \times \frac{4}{\pi} v_{\mathrm{RF}} \sin[(\omega_{\mathrm{RF}} \quad \omega_{\mathrm{LO}}) t] \end{aligned} \tag{8.24}$$

where v_{RF} is the output voltage obtained without the switching (i.e., for a differential amplifier). This means that because of the frequency translation the amplitude of each mixed frequency component is

$$v_o = \frac{2}{\pi} v_{\mathrm{RF}} = v_{\mathrm{RF_{dB}}} \quad 3.9 \text{ dB} \tag{8.25}$$

Mixing results in a factor of 1/2 or 6 dB, but a square wave has a fundamental that is larger by $4/\pi$ or 2.1 dB than a sign wave of the same amplitude, for a net

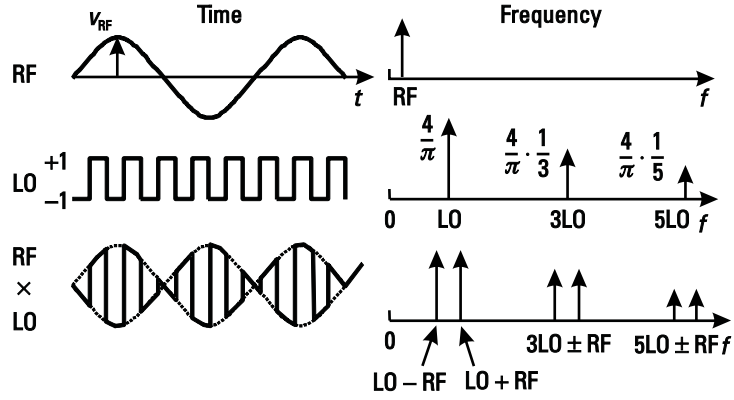

Figure 8.9 Analysis of the switching mixer in the frequency and time domain.

change of 3.9 dB. Third harmonic terms are down by 1/3 or 9.5 dB, while fifth harmonics are 1/5 or 14 dB. Intermodulation (other than mixing between RF and LO) is often due to the RF input and its nonlinearity. Thus, the analysis of the differential pair may be used here. From Chapter 7, the gain v_o/v_i for a differential amplifier with load resistors of R_C and emitter degeneration resistors of R_E per side was given by:

$$v_o = \frac{R_C}{r_e + R_E} v_i \tag{8.26}$$

For a CMOS differential amplifier with no degeneration, the equivalent expression is

$$v_o = g_m R_D v_i \tag{8.27}$$

Thus, a final useful estimate of gain in a bipolar mixer such as the one shown in Figure 8.5 (at one output frequency component) is the following:

$$\frac{v_o}{v_{\text{in}}} = \frac{2}{\pi} \frac{R_C}{r_e + R_E} \tag{8.28}$$

We note that this is voltage gain from the base of the input transistors to the collector of the switching quads. In an actual implementation with matching circuits, these also have to be taken into account.

For the CMOS mixer without degeneration, a useful estimate of the gain at one of the output frequency components is given by:

$$\frac{v_o}{v_{\text{in}}} = \frac{2}{\pi} g_m R_D \tag{8.29}$$

In Figures 8.8 and 8.9, the LO frequency is much greater than the RF frequency (it is easier to draw the time domain waveform). This is upconversion, since the output signal is at a higher frequency than the input signal. Downconversion is shown in Figure 8.10. This is downconversion because the output signal of interest is at a lower frequency than the input signal. The other output component that appears at higher frequency will be removed by the IF filter.

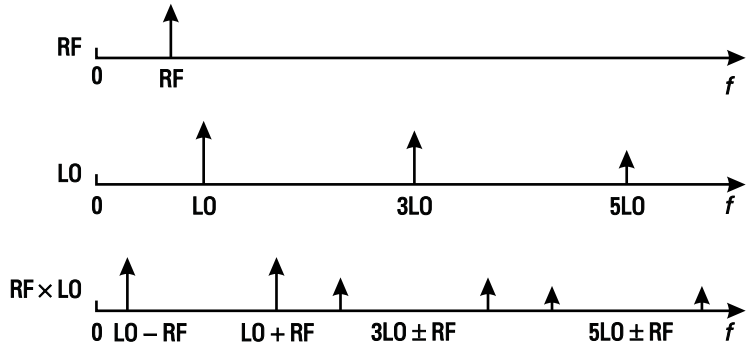

Figure 8.10 Downconversion frequency-domain plot.

Note also that any signals close to the LO or its multiples can mix into the IF. These signals can be other signals at the input, intermodulation between input signals (this tells us we need linear RF inputs), noise in the inputs, or noise in the mixer itself.

8.7 Mixer Noise

Mixer noise figure is somewhat more complicated to define, compared to that of an LNA, because of the frequency translation involved. Therefore, for mixers a slightly modified definition of noise figure is used. Noise factor for a mixer is defined as:

$$F = \frac{N_{o\text{tot}}(\omega_{\text{IF}})}{N_{o(\text{source})}(\omega_{\text{IF}})} \tag{8.30}$$

where $N_{o\text{tot}}(\omega_{\text{IF}})$ is the total output noise at the IF frequency and $N_{o(\text{source})}(\omega_{\text{IF}})$ is the output noise at the IF frequency due to the source. The source and all circuit elements generate noise at all frequencies, and many of these frequencies will produce noise at the output IF frequency due to the mixing action of the circuit. Usually the two dominant frequencies are the input frequency and the image frequency.

To make things even more complicated, *single-sideband* (SSB) noise figure or *double-sideband* (DSB) noise figure are defined. The difference between the two definitions is the value of the denominator in (8.30). In the case of double-sideband noise figure, all the noise due to the source at the output frequency is considered (noise of the source at the input and image frequencies). In the case of single-sideband noise figure, only the noise at the output frequency, due to the source that originated at the RF frequency is considered. Thus, using the double-sideband noise figure definition, even an ideal noiseless mixer with a resistive termination at both the output and image frequencies would have a noise figure of 3 dB. This is because the noise of the source would be doubled in the output due to the mixing of both the RF and image frequency noise to the intermediate frequency. Thus it can be seen that:

$$N_{o(\text{source})\text{DSB}} = N_{o(\text{source})\text{SSB}} + 3\ \text{dB} \tag{8.31}$$

and

$$\text{NF}_{\text{DSB}} = \text{NF}_{\text{SSB}} \quad 3\ \text{dB} \tag{8.32}$$

This is not quite correct, since an input filter will also affect the output noise, but this rule is usually used. Usually single-sideband noise figure is used for a mixer in a superheterodyne radio receiver, since an image-reject filter preceding the mixer removes noise from the image.

Largely because of the added complexity and the presence of noise that is frequency translated, mixers tend to be much noisier than LNAs. The differential pair that forms the bottom of the mixer represents an unattainable lower bound on the noise figure of the mixer itself. Mixer noise will always be higher because

noise sources in the circuit get translated to different frequencies and this often "folds" noise into the output frequency. Generally, mixers have three frequency bands where noise is important:

1. *Noise already present at the IF:* The transistors and resistors in the circuit will generate noise at the IF frequency. Some of this noise will make it to the output and corrupt the signal. For example, the collector resistors will add noise directly at the output IF.
2. *Noise at the RF and image frequency:* Any noise present at the RF and image frequencies will also be mixed down to the IF frequency. For instance, the collector shot noise of Q_1 at the RF frequency and at the image frequency will both appear at the IF frequency at the output.
3. *Noise at multiples of the LO frequency:* Any noise that is near a multiple of the LO frequency can also be mixed down to the IF, just like the noise at the RF.

Also note that noise over a cycle of the LO is not constant, as illustrated in Figure 8.11. At large negative or positive voltage on the LO, dominant noise comes from the bottom transistors. This is the expected behavior, as the LO is causing the upper quad transistors to be switched between cutoff and saturation. In both of these two states, the transistors will add very little noise because they have no gain. We also note that gain from the RF input is maximum when the upper quad transistors are fully switched one way or the other. Thus, a large LO signal that switches rapidly between the two states is ideal to maximize the signal-to-noise ratio.

However, for the finite rise and fall time in the case of a square wave LO signal, or for a sinusoidal LO voltage, the LO voltage will go through zero. During this time, these transistors will be on and in the active region. Thus, in this region they behave like an amplifier and will cause noise in the LO and in the upper quad transistors to be amplified and passed on to the output. As shown in Figure 8.11, for LO voltages near zero, noise due to the upper quad transistors is dominant. At the same time, in this region, gain from the RF is very low, thus for small LO

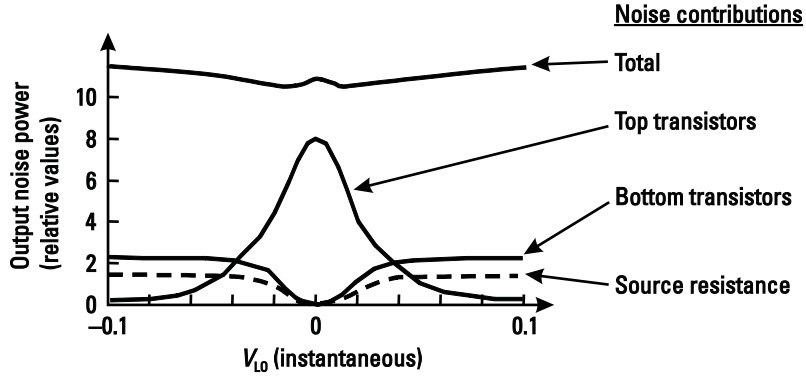

Figure 8.11 Mixer noise shown at various LO levels. Note that not all individual noise contributions are shown.

voltages, the signal-to-noise ratio is very poor, so time spent in this region should be minimized.

In order to determine noise figure, the relative value of the total noise compared to the noise from the source must be determined. Very conveniently, the calculated noise figure is approximately the same when calculated using a slowly swept dc voltage at the LO input or with an actual LO signal (finding the noise figure with the LO voltage held at different dc values and then taking the average). With a dc voltage at the LO terminal, the mixer becomes equivalent to a cascode amplifier and the LO input serves as a gain-controlling signal. When used as a mixer, any noise (or signal) is mixed to two output frequencies, thus reducing the output level. This results in the mixer having less gain than the equivalent differential pair. However, noise from both the RF frequency and the image frequency is mixed to the IF frequency resulting in a doubling of noise power at the output. Thus, the noise prediction based on a swept LO analysis is very close to that predicted using an actual LO signal.

There are a few other points to consider. One is that with a mixer treated as an amplifier, output noise is calculated at f_{RF}, so if some output filtering is included (for example with capacitors across R_C), the noise will be reduced. However, the ratio of total noise to noise due to the source can still be correct. Another point is that in this analysis, noise which has been mixed from higher LO harmonics has not been included. However, as will be shown in Example 8.3, these end up not being very important, so the error is not severe. Some of these issues and points will be illustrated in the next three examples.

Example 8.1: Mixer Noise Figure Determination

For the mixer simulation results shown in Figure 8.11, estimate the mixer noise figure.

Solution:

The noise figure is given by the ratio of total noise to noise from the source. In this example, while individual components of noise from most sources are strongly LO voltage dependent, total noise happens to be roughly independent of instantaneous LO voltage, and the relative value is approximately 11 (arbitrary units). Noise from the source is dependent on instantaneous LO voltage varying from zero to about 1.5 (arbitrary units). This plot illustrates why maximum signal gain and minimum noise figure is realized for a sufficiently large LO signal such that minimal time is spent around zero volts. Thus minimum double-sideband noise figure is

$$\mathrm{NF} = 10\log\frac{N_{o\mathrm{tot}}}{N_{o(\mathrm{source})}} = 10\log\frac{11}{1.5} = 8.65$$

For a real, sinusoidal signal, some time is inevitably spent at zero volts with a resulting time-domain waveform as shown in Figure 8.12.

In the diagram, the effective input noise is reduced down to about 1.1, and as a result, the noise figure is increased to about 10 dB.

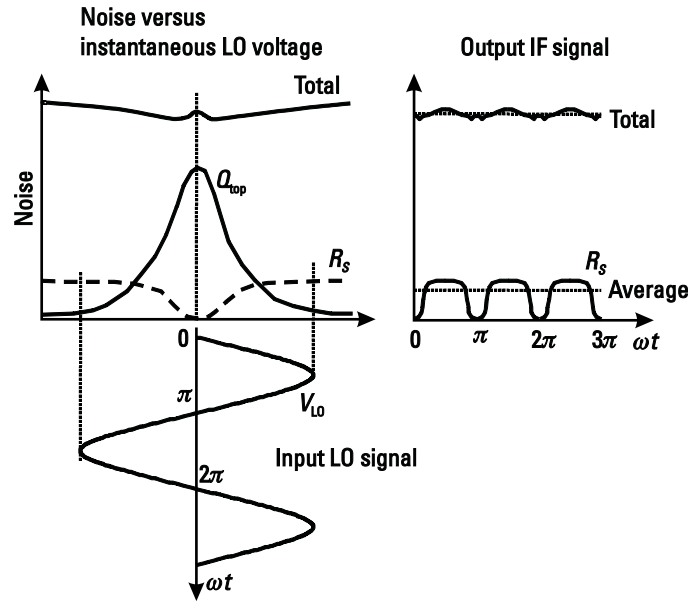

Figure 8.12 Noise calculations.

Example 8.2: Mixer Noise and Gain with Degeneration

In this example for a mixer as in Figure 8.13, some equations and simulation results will be shown for noise and gain versus degeneration.

Without consideration of input and output matching, the noise sources associated with R_S and R_E both have gain approximately equal to the signal gain given by:

$$\frac{v_{\text{no},R_E}}{v_{n,R_E}} = \frac{v_{\text{no},R_S}}{v_{n,R_S}} = \frac{v_o}{v_{\text{RF}}} \quad \frac{2}{\pi}\frac{R_C}{r_e + R_E}$$

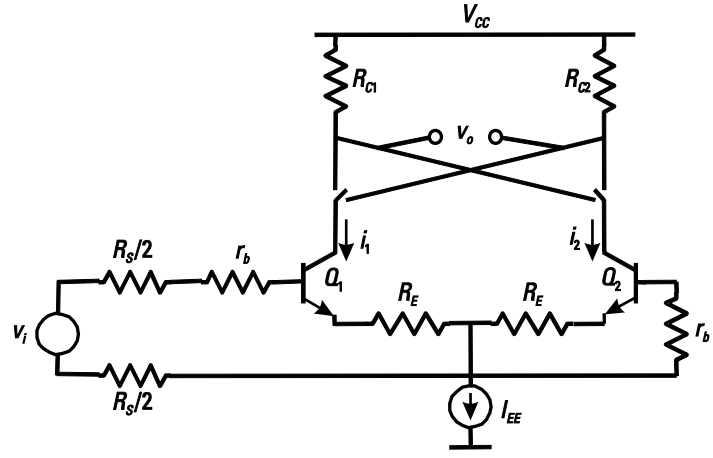

Figure 8.13 Mixer with switching.

where v_{no,R_E} and v_{no,R_S} are the output noise due to R_S and R_E and v_{n,R_E} and v_{n,R_S} are the noise voltage associated with each resistor. Thus, with degeneration, gain will decrease unless R_C is increased to match the increase of R_E. Similarly, the noise figure will degrade with increasing R_E, since the noise due to R_E is given by:

$$v_{n,R_E} = \sqrt{4kTR_EB}$$

Simulation results are shown in Figure 8.14. It can be seen that gain decreases rapidly for increased values of R_E, from 12 dB to about 0 dB for R_E from 0 to 100 . Noise figure increases by about 4 dB over the same change of R_E. We note that matching will make a difference and adding degeneration resistance changes the input impedance, which will indirectly change the noise due to the effect of input impedance on base shot noise.

We also note that, as predicted by theory, gain is about 4 dB lower for a mixer compared to a differential pair. Noise is higher for the mixer by about 3 to 5 dB due to noise mixing from other frequencies and noise from the switching quad.

Example 8.3: Mixer Noise Components—Sources and Frequencies

Noise in a mixer, such as shown in Figure 8.15, comes from a variety of sources and is mixed to the output from a variety of frequencies. In this example, we will show the typical relative levels of these noise components. Many simulators could provide the information for this analysis, but instead of discussing the simulation, here the results will be discussed.

Table 8.1 shows the noise from various sources and from various frequencies. Noise has been expressed as a voltage or as a squared voltage instead of a density by assuming a bandwidth of 1 MHz. Noise from the bottom components (RF input transistors, current sources, input resistors, and bias resistors) has been further broken down in Table 8.2. Noise from the source resistor has been shown in brackets, as its effect is also included with total bottom noise.

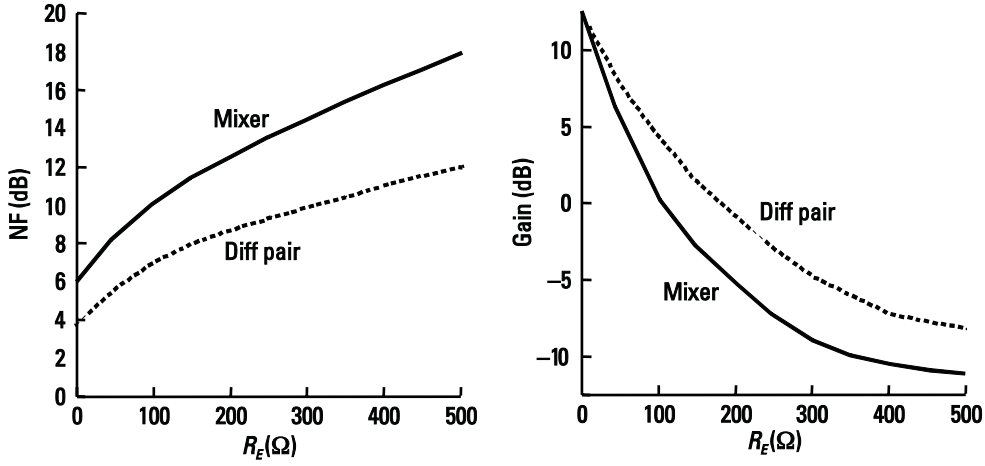

Figure 8.14 Noise simulation results.

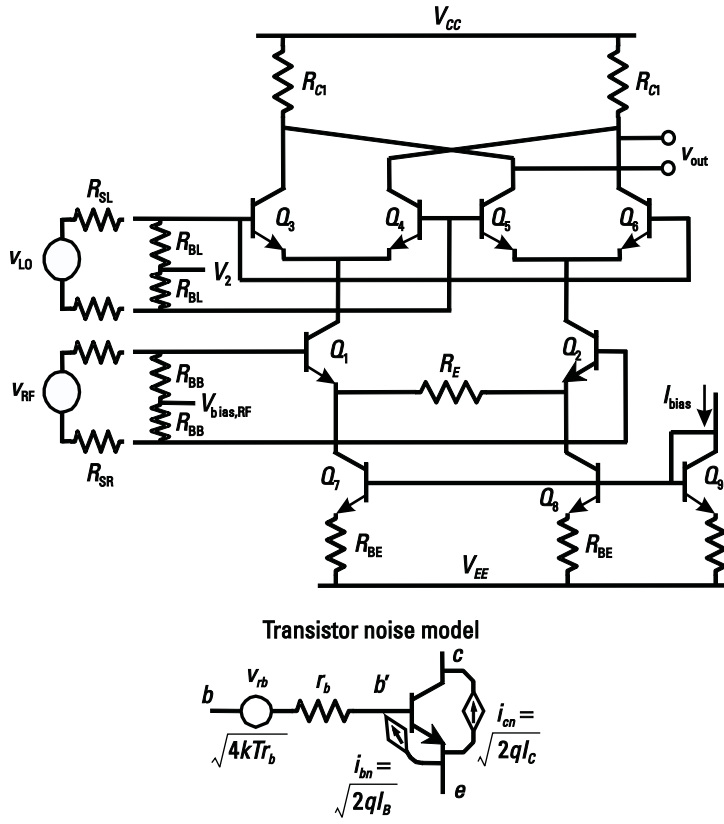

Figure 8.15 Mixer for noise analysis and transistor noise model.

The final row in Table 8.1 shows for each specified input frequency, the total number of equivalent frequency bands that have approximately the same noise. For example, for every noise input at the RF frequency, there is an approximately equal input at the image frequency. Thus, for an input at the RF frequency, or around other harmonics of the LO frequency, the multiplier is 2, while for noise inputs at the IF frequency, the multiplier is 1. Different frequencies are dominant for different parts of the mixer. For the bottom circuitry, the most important input frequency is

Table 8.1 Mixer Noise and Dominant Sources of Noise

	Differential Output Voltage at ω_{IF}				*Total*
Source	*Input at ω_{IF}*	*Input at ω_{RF}*	*Input at $2\omega_{LO}$-ω_{IF}*	*Input at $3\omega_{LO}$-ω_{IF}*	*$10^{-6}V^2$*
(R_{SR})	(0.7 nV)	(0.72 mV)	(0.2 nV)	(0.11 mV)	(1.08)
Bottom	0.9 nV	1.77 mV	4.0 nV	0.29 mV	6.44
v_{rbquad}	0.25 mV	0.03 mV	0.20 mV	0.02 mV	0.15
i_{cquad}	0.26 mV	0.15 mV	1.5 μV	0.03 mV	1.17
i_{bquad}	36 nV	0.13 mV	25 nV	0.03 mV	0.04
v_{R_L}	1.77 mV	16 nV	4.4 μV	1.3 nV	3.13
Frequency bands	1	2	2	2	10.93

Table 8.2 Breakdown of Noise in the RF Stage

Source	*Noise Out* V^2	*Source*	*Noise Out* V^2
$r_{b,Q1,2}$	0.51 10^{-6}	R_{SR}	1.08 10^{-6}
$r_{e,Q1,2}$	0.08 10^{-6}	R_{BB}	1.08 10^{-6}
i_b	0.06 10^{-6}	R_{BE}	0.20 10^{-6}
i_c	1.21 10^{-6}	R_E	2.22 10^{-6}
Total Q1,2	1.86 10^{-6}	Total	6.44 10^{-6}

the RF (and image frequency). For the load resistor, the only important component occurs at the IF, while for the quad switching transistors, both the RF and the IF are important. Generally, noise around the LO second harmonic or higher harmonics is not important; however, the third harmonic will add a bit of noise from the bottom circuitry.

As for sources of noise, the bottom circuitry is seen to be the dominant factor (true only if the LO amplitude is large enough to switch the quad transistors fully). Thus, to minimize noise, the RF input stage must be optimized, similar to that of an LNA design. Of the bottom noise sources, degeneration can quickly become the dominant noise source for a high linearity design. If linearity allows, it is possible to use inductor degeneration for optimal noise and power matching. As well, in this example, bias resistors contributed significantly to the noise. In an inductively degenerated mixer, bias resistors can be made significantly larger to minimize the noise contribution from them.

Double-sideband noise figure can be calculated as follows:

$$\text{DSB noise figure} = 10\log_{10}\left(\frac{10.93}{1.08}\right) = 10.1\ \text{dB}$$

Minimum noise figure from the RF stage can be calculated as:

$$\text{DSB NF}_{\min} = 10\log_{10}\left(\frac{6.44}{1.08}\right) = 7.75$$

Note that a single sideband noise figure would consider source noise in the RF band only, thus noise figure would have been higher by about 3 dB.

8.7.1 Summary of Bipolar Mixer Noise Components

Noise equations suitable for hand calculation of noise figure can be determined by considering each noise source and its transfer function to a common point. Total noise compared to noise from the source is the noise factor. For these equations, there is a single-source resistance of value R_S. Noise from the source can be referred to the output by

$$\overline{v}_{noR_S}^2 = 1 \quad 2 \quad 4kTR_SA^2 \tag{8.33}$$

By way of brief explanation, the noise source of 4 kTR_S is multiplied by A^2, the square of the voltage gain. It is assumed here that the bias resistors are large rela-

tive to R_S. If R_B is comparable to R_S, then a voltage divider would have to be added to the equation. The extra factors of 1 and 2 in this and the following equations indicate the number of sources (in this case a single source resistance with thermal noise) and the number of bands (in this case the RF and the image band). The voltage gain A is given by

$$A = \frac{2}{\pi} \frac{R_{CL}}{r_e + R_E} \tag{8.34}$$

It is assumed that the mixer input impedance is high compared to the source resistance. The factor R_{CL}, the equivalent load resistance per side, is the parallel combination of collector resistor R_C and extra load resistor R_L and R_E is the added degeneration resistor per side.

The total output noise due to collector shot noise is

$$\overline{v}_{noI_c}^2 = 2 \quad 2 \quad 2kTg_m \quad \frac{2}{\pi} R_{CL} \div \quad ^2 = 2 \quad 2 \quad 2kTg_m A^2 (r_e + R_E)^2 \tag{8.35}$$

The factor of $2/\pi$ is due to the switching in the upper quad, one factor of two is due to the fact that there are two devices, and another factor of two to account for the image band.

Noise at the output due to the base resistance is given by

$$\overline{v}_{nor_b}^2 = 2 \quad 2 \quad 4kTr_b A^2 \tag{8.36}$$

Noise at the output due to emitter degeneration R_E is given by

$$\overline{v}_{noR_E}^2 = 2 \quad 2 \quad 4kTR_E A^2 \tag{8.37}$$

As indicated above, this assumes that there is a resistor of value R_E on each side. Finally, the noise due to the load resistance R_{CL} is given by

$$\overline{v}_{noR_{CL}}^2 = 2 \quad 1 \quad 4kTR_{CL} \tag{8.38}$$

This assumes that there is a collector resistor R_C and an extra load resistor R_L per side, both contributing to the noise and this is shown by the factor of 2. At the output, noise mostly occurs directly in the IF band, hence there is only one band to consider. If R_L were considered noise free, only R_C would contribute to the noise, and this noise would also have undergone a voltage division. The resulting noise would then be

$$\overline{v}_{noR_C}^2 = 2 \quad 1 \quad 4kTR_C \quad \frac{R_L}{R_L + R_C} \div \quad ^2 \tag{8.39}$$

Noise factor is found by combining all these equations as

$$F = \frac{N_{o,tot}}{N_{o,src}} = 1 + \frac{g_m (r_e + R_E)^2}{R_S} + \frac{2r_b}{R_S} + \frac{2R_E}{R_S} + \frac{R_{CL}}{R_S A^2} \tag{8.40}$$

As a reminder, as indicated above, for this equation, it is assumed that there are two each of R_C, R_L, and R_E, and it is assumed that both R_C and R_L contribute to the noise. This equation is for double-sideband noise figure. For single-sideband noise factor, input noise in the denominator of (8.40) would be assumed to be due only to the noise in the RF band, thus single-sideband noise factor would be 3 dB higher than double-sideband noise factor.

8.7.2 Summary of CMOS Mixer Noise Components

In this section, noise calculations similar to those in the previous section will be shown for a CMOS mixer without degeneration. Noise from the source can be referred to the output by

$$\overline{v}_{noR_S}^2 = 1 \quad 2 \quad 4kTR_SA^2 \tag{8.41}$$

Again, the factor of 1 is because there is a single source resistance and the factor of 2 indicates that noise is coming from two bands (the RF and the image band). Voltage gain A for the CMOS mixer is given by

$$A = \frac{2}{\pi} g_m R_{DL} \tag{8.42}$$

where R_{DL}, the equivalent load resistance, is the parallel combination of drain resistor R_D per side and extra load resistor R_L per side. The total output noise due to drain noise is

$$\overline{v}_{noI_c}^2 = 2 \quad 2 \quad 4kT\gamma g_m \quad \frac{2}{\pi} R_{DL} \div \quad ^2 = 2 \quad 2 \quad \frac{4kT\gamma}{g_m} A^2 \tag{8.43}$$

The factor of $2/\pi$ is due to the switching in the upper quad.

Noise at the output due to the gate resistance is given by

$$\overline{v}_{nor_g}^2 = 2 \quad 2 \quad 4kTr_gA^2 \tag{8.44}$$

Finally, the noise due to the load resistance R_{DL} is given by

$$\overline{v}_{noR_{DL}}^2 = 2 \quad 1 \quad 4kTR_{DL} \tag{8.45}$$

As for the bipolar case, this assumes that there is a drain resistor R_D and an extra load resistor R_L per side, both contributing to the noise and this is shown by the factor of 2, and that noise contribution is only from the single IF band.

Noise factor for the CMOS mixer is found by combining all these equations as

$$F = \frac{N_{o,tot}}{N_{o,src}} = 1 + \frac{2\gamma}{g_m R_S} + \frac{2r_g}{R_S} + \frac{R_{DL}}{R_S A^2} \tag{8.46}$$

As a reminder, as indicated above, for this equation, it is assumed that there are two each of R_C and R_L, and it is assumed that both R_C and R_L contribute to the

noise. This equation is for double-sideband noise figure. For single-sideband noise factor, input noise in the denominator of (8.46) would be assumed to be due only to the noise in the RF band; thus, single-sideband noise factor would be 3 dB higher than double-sideband noise factor.

8.8 Linearity

Mixers have both desired and undesired nonlinearity. The mixing action of the switching quad is what is necessary for the operation of the circuit. However, mixers also contain amplifiers that can be nonlinear. Just as in Chapter 7, these amplifiers have linearity requirements.

8.8.1 Desired Nonlinearity

A mixer is inherently a nonlinear device. Linear components have output frequencies equal to the input frequencies, and so no mixing action can take place for purely linear inputs. This *desired nonlinearity* comes from the switching action of the quad transistors as determined by the nonlinear characteristics of the transistor. Thus, if we have two tones at the input, the desired outputs for a switching mixer are as shown in Figure 8.16.

The only desired output may be at the IF, but the other components are far enough away they can easily be filtered out.

8.8.2 Undesired Nonlinearity

Undesired nonlinearity can occur in several places. One is at the RF input, which converts the input signal into currents i_1 and i_2 (see Figure 8.13). The reason for adding degeneration resistors R_E, or for sizing CMOS input transistors appropriately, is to keep this conversion linear, just as in the case of an LNA. However, some nonlinearity will still be present. The resulting output is shown in Figure 8.17.

Thus, the only difference in this circuit compared to the differential pair being considered in Chapter 7 is that now there is a frequency translation. Thus, just as before, for a bipolar mixer with a differential pair without degeneration, the input

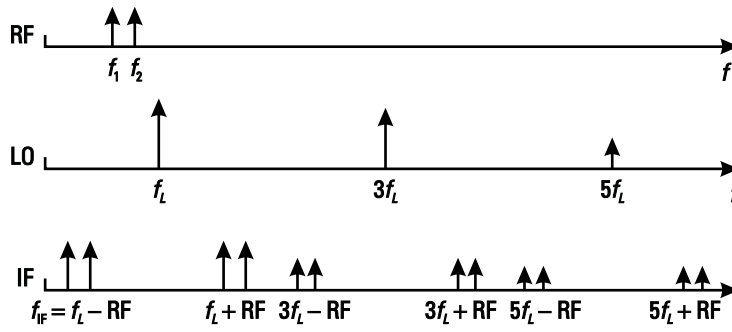

Figure 8.16 Mixer expected outputs in the frequency domain.

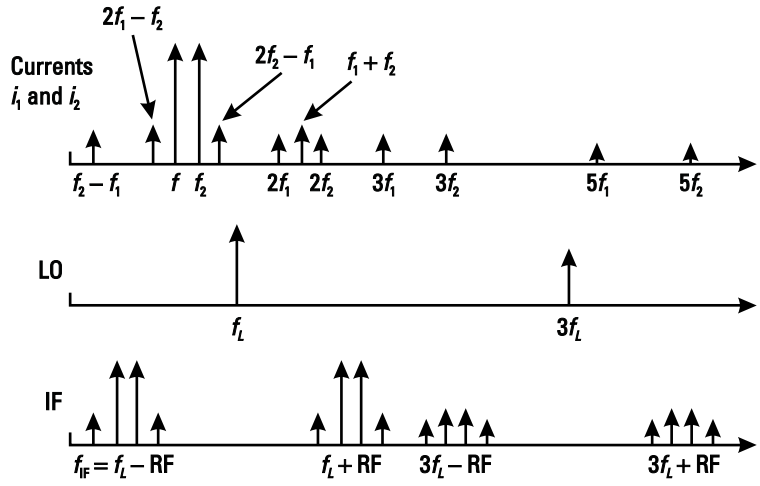

Figure 8.17 Mixer with nonlinearity in the RF input stage.

IP3 is related to v_L the linear extrapolation of the axis of the transfer curve. The resulting expression is

$$v_{IP3} = 4v_T = 2v_L \tag{8.47}$$

As discussed in Chapter 7, for small values of degeneration resistor, the same value of $2v_L$ is used to estimate linearity resulting in the following

$$v_{IP3} = 2v_L = 4v_T + 2I_{EE}R_E = 2I_{EE}(r_e + R_{EE}) \tag{8.48}$$

With larger values of resistor, when R_{EE} is significantly larger than r_e, an estimate for the linear voltage extension v_L is often assumed to be roughly equivalent to a 1-dB compression point; thus, IP3 is roughly three times the voltage resulting in input IP3 being estimated as

$$v_{IP3} = 3v_L = 3I_{EE}(2r_e + R_{EE}) \tag{8.49}$$

Note that R_{EE} is the total resistance between the two emitters and that I_{EE} is the total current being supplied to the differential pair. In Figure 8.5, the total current supplied I_{EE} is $2I_o$ and the total resistance R_{EE} is $2R_E$ so appropriate adjustments need to be made.

For a CMOS mixer based on a CMOS differential pair, from Chapter 7 linearity is given by

$$v_{IP3} = 3.266v_L = 3.266\frac{I_{EE}}{g_m} \tag{8.50}$$

Thus for both bipolar with degeneration and with CMOS mixers the 1-dB compression point is roughly equal to v_L the linear extrapolation of the small signal transfer curve. We note that v_L is roughly equal to the voltage for which visible clipping would be seen, and this estimate for 1-dB compression voltage can also be used in other parts of the mixer.

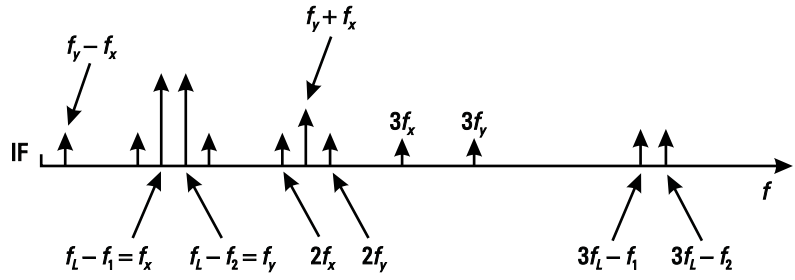

Figure 8.18 Mixer with nonlinearity in switching quad.

With nonlinearity in the RF input stage, the currents i_1 and i_2 are composed of a large number of frequency components. Each of these frequencies is then mixed with each of the LO harmonics, producing the IF output. Many of the IF frequencies (such as mixing of harmonics of the radio frequencies with harmonics of the LO) have not been shown.

Note that if the RF input were perfectly linear, mixing action would proceed cleanly with the result as previously shown in Figure 8.16. However, undesirable frequency components can be generated because of nonlinearity in the output stage, for example, due to limiting action. In such a case, the two IF tones at $f_L - f_1$ and $f_L - f_2$ would intermodulate, producing components at $f_1 - (2f_1 - f_2)$ and $f_L - (2f_2 - f_1)$ in addition to harmonics of $f_L - f_1$ and $f_L - f_2$. This is shown in Figure 8.18. Again, a 1-dB compression point can be estimated as the voltage for where clipping just starts to occur.

In some cases, the saturation of the switching quad may be the limiting factor in the linearity of the mixer. For example, if the bias voltage on the base of the quadrature switching transistors v_{quad} is 2V and the power supply is 3V, then the output can swing from about 3V down to about 1.5V, for a total 1.5V peak swing. If driving 50 , this is 13.5 dBm. For larger swings, a tuned circuit load can be used as shown in Figure 8.19. Then, the output is nominally at V_{CC} with equal swing above and below V_{CC}, for a new swing of $3V_p$ or $6V_{p\text{-}p}$ which translates to 18.5 dBm into 50 . We note that an on-chip tuned circuit may be difficult to realize for a downconverting mixer, since the output frequency is low, and therefore the inductance needs to be large.

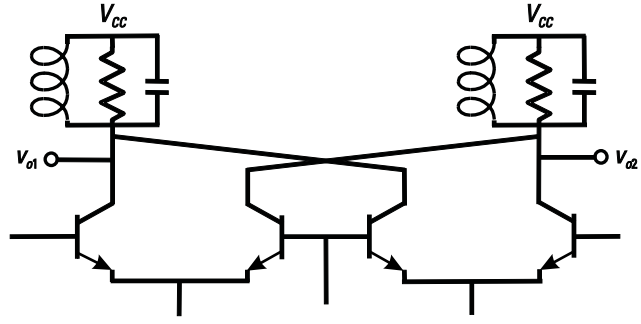

Figure 8.19 Tuned load on a mixer.

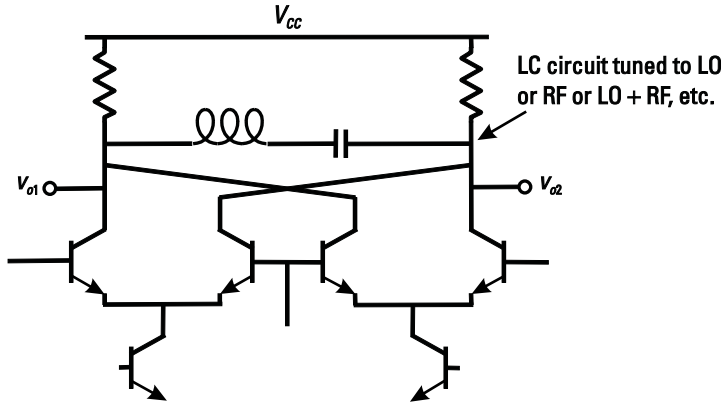

Figure 8.20 Series LC between mixer outputs.

If the output needs to drive a low impedance such as 50 , often an emitter follower is used at the output. This can be a fairly broadband circuit, since no tuned components need be used, but could cause some distortion.

8.9 Improving Isolation

It is also possible to place an inductor-capacitor (LC) series circuit across the outputs, as shown in Figure 8.20, to reduce LO or RF feedthrough, or to get rid of some upconverted component. This can also be accomplished with the use of a lowpass filter by placing a capacitor in parallel with the load, as shown in Figure 8.21, or using a tuned load as shown in Figure 8.19.

8.10 General Design Comments

So far in this chapter, we have discussed basic theory of the operation of mixers. Here we will provide a summary of some general design guidelines to help with the trade-offs of optimizing a mixer for a particular application.

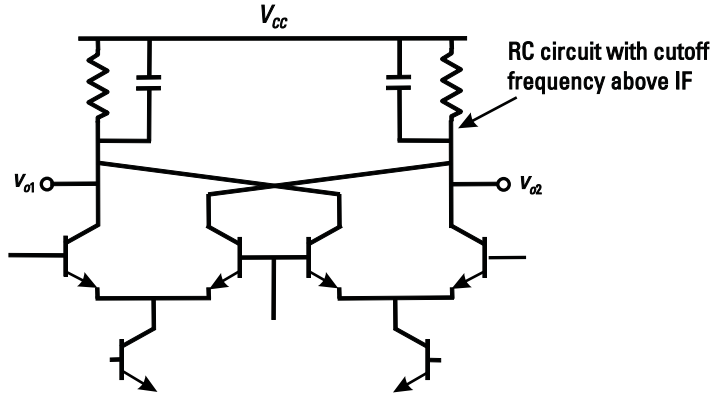

Figure 8.21 Parallel RC circuit across the output.

8.10.1 Sizing Transistors

The differential pair that usually forms the bottom of a double-balanced mixer is basically an LNA stage, except that the transistors are usually optimized first for linearity and gain, then for noise. The switching quad transistors are the part of the circuit unique to the mixer. Usually, in a bipolar design the switching transistors are sized and biased so that they operate close to their peak f_T. If the differential pair transistors are biased at their minimum noise current, then the switching transistors end up being about one-eighth the size. In a typical low voltage CMOS process, the switching quad transistors are often sized according to the available input swing and voltage headroom and this often means operating at a lower current density than would be best for maximum f_T. For maximum speed, of course operation close to optimal f_T is desirable; however, this requires a V_{GS} approaching that of the rail voltage.

8.10.2 Increasing Gain

As shown previously in (8.28), without matching considerations, and assuming full switching of the upper quad, voltage gain in a bipolar mixer is estimated by:

$$\frac{v_o}{v_{in}} = \frac{2}{\pi}\frac{R_C}{r_e + R_E} \tag{8.51}$$

To increase the gain, the choices are to increase the load resistance R_C, to reduce degeneration resistance R_E, or to increase the bias current I_o. Since the output bias voltage is approximately equal to $V_C \quad V_{CC} - I_o R_C$, increased gain will be possible only if adjustments to R_C, R_E, or I_o do not cause the switching transistors to become saturated. In a CMOS mixer without degeneration, the gain is of the form:

$$\frac{v_o}{v_{in}} = \frac{2}{\pi} g_m R_D \tag{8.52}$$

Gain can be increased by increasing R_D or by increasing g_m by changing the transistor size or bias current. Assuming dc current is flowing through R_D, an increase of current or in the size of R_D directly results in reduced headroom. Increasing g_m by increasing transistor size will mean less voltage drop across the transistor, which is good for headroom, but it makes the transistor itself less linear since linearity is related to the overdrive voltage, $V_{GS} - v_T$. Note that with degeneration, the gain equation takes on a form similar to that of the bipolar mixer and with similar considerations for increasing gain.

8.10.3 Improvement of IP3

How to increase IP3 depends on which part of the circuit is compressing. Compression can be due to overdriving of the lower differential pair, clipping at the output, or the LO bias voltage being too low causing clipping at the output of the bottom differential pair. After a problem has been identified, IP3 can be improved by one or more of the following:

1. If the compression is due to the bottom differential pair (RF input), then in a bipolar mixer, or in a CMOS mixer with degeneration, linearity can be improved by increasing R_E or by increasing bias current. We note from (8.51) that increased R_E will result in decreased gain. Increased bias current will increase the gain slightly through a reduction of r_e although this effect will be small if degeneration resistance is significantly larger than r_e. In a CMOS mixer without degeneration, linearity can be improved by increasing the current density of the transistor, by reducing the transistor size or increasing the current.
2. Compression caused by clipping at the output is typically due to the quad transistors leaving the proper operating region (going into saturation for bipolar, leaving saturation for CMOS transistors). This can be avoided by moving the output bias level to a higher voltage by reducing the bias current or reducing the load resistors. However this will reduce the gain. In a CMOS mixer, another option is to increase the size of the switching transistors. Another possibility, although not usually a practical one, is to increase V_{CC}. The use of a tuned output circuit is effectively equivalent to raising the supply voltage.
3. If compression is caused by clipping at the collector of the RF input differential pair, then increasing the LO bias voltage will improve linearity; however, this may result in clipping at the output.

It is possible to conduct a series of tests on the actual circuit or in a circuit simulator to determine where the compression is coming from. First, at least in simulation, it is possible to increase the power supply voltage to some higher value. If the compression point is increased, then output clipping is a problem. If the compression is unchanged, then the problem is not at the output. Next, one can determine if the LO bias voltage is sufficiently high by increasing it further. If linearity is not improved, then this was not the cause of the original linearity problem. Then, having eliminated output clipping or LO biasing problems, one can concentrate on the lower differential pair. As discussed in the previous paragraphs, its linearity can be improved by increasing current or by increasing R_E or in a CMOS mixer by reducing the transistor size.

8.10.4 Improving Noise Figure

Noise figure will be largely determined by the choice of topology with the opportunity for the lowest noise provided by the simultaneous matched design of Section 8.12.3. The next most important factor is the value of the emitter degeneration resistor, or the size of the input transistors in a CMOS mixer. To minimize noise, the emitter degeneration resistor should be kept as small as possible; however, with less degeneration, getting the required linearity will require more current. Similarly, for a CMOS mixer, higher g_m in the input transistors will generally improve noise but higher current is required to maintain linearity.

8.10.5 Effect of Bond Pads and the Package

In a single-ended circuit, such as an LNA, the effect of the bond pads and the package is particularly important for the emitter, since this is a low impedance node and

has a strong influence on the gain and noise. For a differential circuit, such as the mixer, the ground is a virtual ground and the connection to the external ground is typically through a current source. Thus, the bond pad on the ground node here has minimal impact on gain and noise. At the other nodes, such as inputs and outputs, the bond pads will have some effect since they add a series inductor. However, this can be incorporated as part of the matching network.

8.10.6 Matching, Bias Resistors, Gain

If the base of the RF transistors were biased using a voltage divider with an equivalent resistance of 50 , the input would be matched over a broad band. However, the gain would have dropped by about 6 dB compared to the gain achievable when matching the input reactively with an LC network or with a transformer. For a resistively degenerated mixer, the RF input impedance (at the base of the input transistors) will be fairly high, for example, with $R_E = 100$, Z_{in} can be of the order of a kilohm. In such a case, a few hundred ohms can make it easier to match the input; however, there will be some signal attenuation and noise implications.

At the output, if matched, the load resistor R_o is equal to the collector resistor R_C and the voltage gain is modified by a factor of 0.5. Furthermore, to convert from voltage gain A_v to power gain P_o/P_i, one must consider the input resistance R_i and load resistance $R_o = R_C$ as follows:

$$\frac{P_o}{P_i} = \frac{\dfrac{v_o^2}{R_o}}{\dfrac{v_i^2}{R_i}} = \frac{v_o^2}{v_i^2}\frac{R_i}{R_o} = A_v^2\frac{R_i}{R_o} \qquad \frac{2}{\pi}\frac{R_C/2}{R_E} \div \frac{R_i}{R_C} \tag{8.53}$$

Note the equation is also valid for CMOS mixers if a drain resistor R_{D} is used instead of the collector resistor R_C.

Example 8.4: CMOS Mixer Design

Design a CMOS mixer without degeneration in a commercial 0.13-μm CMOS process that will take a 5.2-GHz input and convert it to 200 MHz. Design for an input referred v_{IP3} of 250 mV, double-sideband noise figure of 10 dB, and voltage conversion gain of 6 dB.

Solution:

Curves for a 20-μm transistor with minimum channel length for current and transconductance are given in Figure 8.22(a). Figure 8.22(b) shows the I_o/g_m ratio that will be used to determine linearity, noting that I_o is twice the nominal bias current for each transistor. The curves are also compared to best-fitting square law equations showing that reasonable results are obtained over some range of v_{GS}. These curves will be scaled to achieve the desired final results. Note that if transistor size and current are both scaled, the operating voltage (V_{GS}) and linearity will roughly remain the same; however, g_m will scale with current and size. Thus, these curves

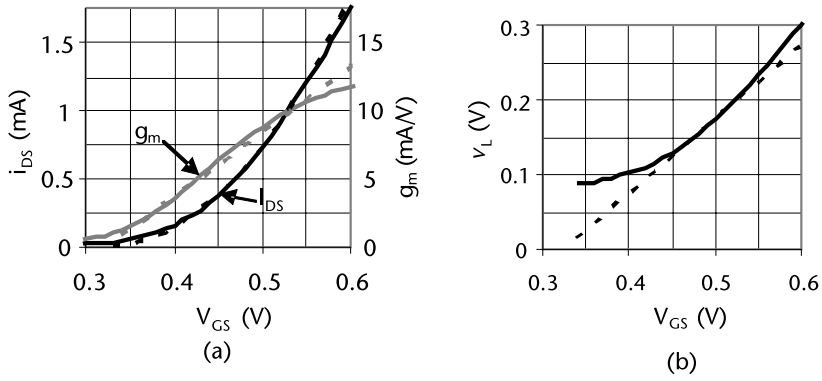

Figure 8.22 Simulated (solid) and predicted (dashed) characteristics for a 20-μm transistor: (a) current i_{DS} and transconductance g_m where predicted i_{DS} is $24(V_{GS} - 0.325)^2$ and predicted g_m is $48(V_{GS} - 0.325)$ and (b) v_L or 2 i_{DS}/g_m.

can be used to achieve the desired linearity, then scaling is done to achieve the desired g_m, for example to achieve the desired gain and noise figure.

Linearity requires not only that the input differential pair has the appropriate linearity as given by the above plots, but that the bias voltages are chosen appropriately to allow sufficient signal swing at each of the nodes. Since the standard cross-coupled mixer has three transistors stacked, plus a load resistor, biases must be chosen carefully. As shown in Figure 8.23, a transistor must be biased for proper dc operation and also to be able to handle the ac signal v_{swing} such that

$$V_{DS} \quad V_{GS} \quad V_T + v_{swing}$$

Given a linearity number, a good practice is to design the circuit such that it can handle the swing at the 1-dB compression point. Thus if v_{IP3} is 250 mV, the 1-tone, 1-dB compression point is 9.6 dB lower or about 83 mV. For a bit of a safety factor, the design will be done for 100-mV peak differential, (or 50 mV per side). For a voltage gain of 2 (6 dB) the voltage swing on each output node is 100 mV while on the input it is 50 mV. For the required input signal level, from the simulated curve in Figure 8.22(b), 0.4 V for V_{GS} would appear to be adequate, but a larger V_{GS} will result in more linearity. As well, linearity at frequency is not perfectly predicted by the dc transfer curves, so ultimately simulations will need to be done including extracted parasitic components. From the curve, biasing the transistor at about

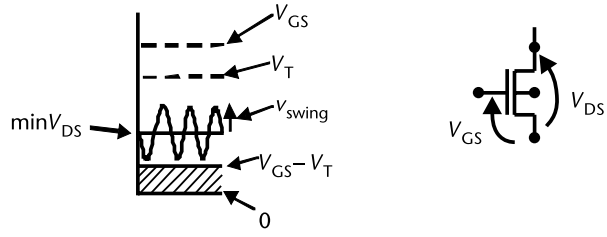

Figure 8.23 Biasing a transistor to leave room for output signal swing.

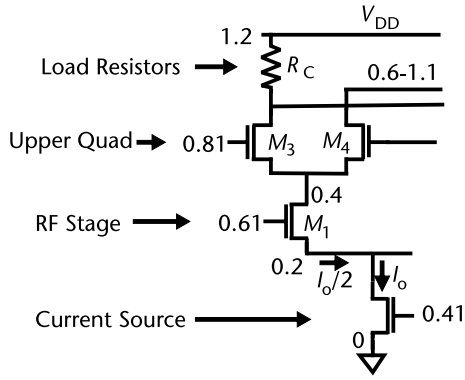

Figure 8.24 Biasing a mixer with a V_{GS} of 0.41V and a V_{DS} of 0.2V.

0.41V would result in a bias current of 0.2 mA through each transistor in the RF stage, which means I_o is 0.4 mA. For this low level, a V_{DS} of 0.2V is adequate. Bias voltages are shown in Figure 8.24. This leaves a significant amount of head room for the output signals and also allows for a significantly larger LO signal amplitude (e.g., of the order of 600-mV peak as required for good switching performance). We note that due to the body effect, threshold voltage is increased, hence all bias voltages will have to be increased accordingly.

The next step is to set the gain. At 0.2 mA, g_m is about 4 mA/V. To achieve a gain of 2, from (8.42) load resistance can be calculated to be

$$R_{DL} = \frac{A \times \pi}{2g_m} = \frac{2\pi}{2 \quad 0.004} = 785$$

A value of 1 k is selected allowing for some additional loss in the circuit. At 0.2 mA, the bias drop across 1 k is 0.2V, which should be sufficient. Calculating the noise using (8.46) including drain current and the load resistor, results in

$$F = 1 + \frac{1}{A_v^2}\frac{R_{CL}}{R_s} + \frac{2\gamma}{g_m R_s} = 1 + \frac{1}{2^2}\frac{1{,}000}{50} + \frac{2 \quad 1}{4m \times 50} = 1 + 5 + 10 = 16 \qquad NF = 12.0 \text{ dB}$$

This clearly does not meet the specifications so adjustments will be needed. We note that scaling both the transistor size and the current by a factor of k will increase g_m and I_{DS} by a factor of k. To keep A_V the same, R_{CL} must be scaled by a factor of $1/k$. The result is that linearity, determined by the ratio of current to transconductance remains the same, but the second and third terms of the noise factor terms above are both scaled by $1/k$. To achieve 10 dB, F must be 10, or added noise factor is 9. However, this assumes that bias resistors are large, and that noise due to the gate resistance and the cascade transistors can be neglected. With good design (e.g., using multiple fingers and large bias resistors), the assumption is reasonable; however, as seen earlier, the cascade noise may still be significant. As well, 1/f noise can be significant even at a few hundred megahertz. Careful simulations are needed to determine how significant these are, but for this example, we will scale the transistor and current up by a factor of 2.5, hence the second and third term will scale

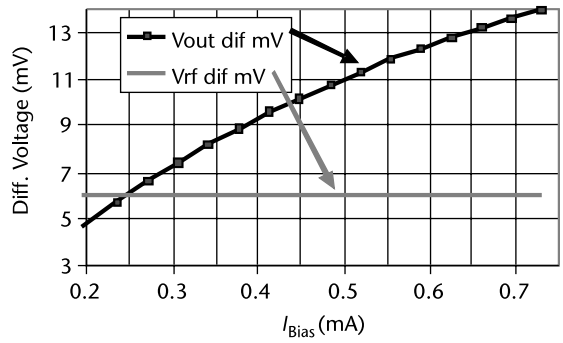

Figure 8.25 CMOS mixer voltage levels versus bias current.

down by a factor of 0.4 and final noise factor will be 7 for a noise figure of about 8.5 dB, leaving some margin for additional sources of noise. The resulting transistor size is 50 μm, and the bias current is 0.5 mA per side for a total of 1 mA.

The initial design is now complete and simulations are performed to verify the performance. First the LO available input power was set to 0 dBm. For a 50 load, this would be equivalent to 316-mV peak, but because impedance was not matched to 50 , the actual LO input voltage was 615-mV peak. Then the bias current was swept to verify the gain calculation and the results are shown in Figure 8.25. At 0.5 mA per side, for an RF input of about 6 mV, the output was at 11 mV, slightly lower than predicted. A gain of 2 could have been achieved by increasing the current to about 0.57 mA; however, further simulations were all done at 0.5 mA.

Then fixing the bias current to 0.5 mA, the LO available power level was swept from 10 to 10 dBm and it was observed that a maximum output voltage of just over 11 mV and a minimum noise figure of just under 10 dB occurred for an LO at 2.5 dBm as seen in Figure 8.26. Thus, 0 dBm is very close to the optimal and this value was used for characterization. Although the gain is a bit lower than spec, the noise is nearly exactly at the desired value. The noise summary indicated that the noise due to the drain channel noise and the load resistor resulted in a noise factor of 7.1 nearly exactly as predicted. From the summary, the main additional noise was due to the upper quad transistors.

Figure 8.27 shows mixer linearity. This would normally be plotted with the *y*-axis in dBm, however, because the impedances are not matched, it has instead

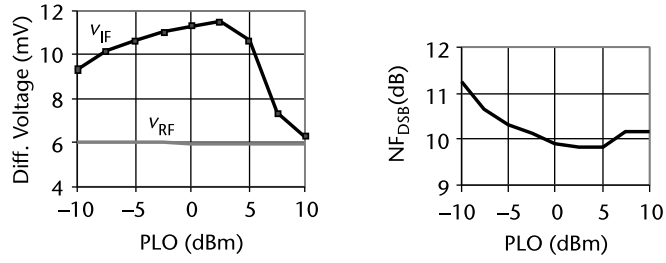

Figure 8.26 CMOS mixer voltage levels and noise versus LO available power.

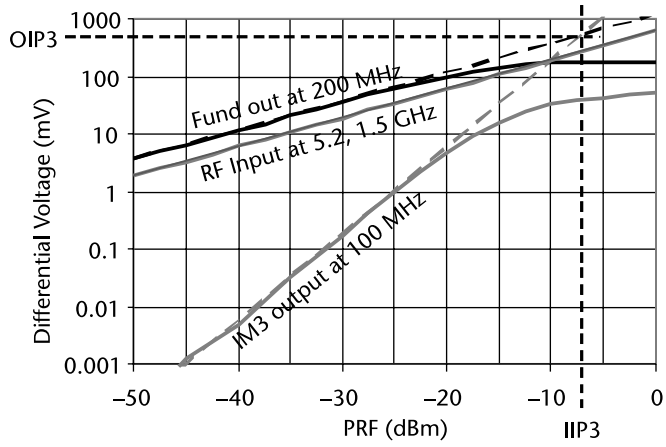

Figure 8.27 CMOS mixer linearity.

been plotted in millivolts, and RF input and IF output voltages have been shown. IP3 referred to the available power from the RF port is at about –7.1 dBm. Using markers on the original RF input plot (not shown), this corresponds to a voltage level of 265 mV, slightly better than the specified linearity. The extrapolated output and the extrapolated third-order output are both at 510 mV.

8.11 Image-Reject and Single-Sideband Mixer

Mixing action as shown in Figure 8.28 always produces two sidebands: one at $\omega_1 + \omega_2$ and one at $\omega_1 - \omega_2$ by multiplying $\cos \omega_1 t \quad \cos \omega_2 t$. It is possible to use a filter *after* the mixer in the transmitter to get rid of the unwanted sideband for the upconversion case. Similarly, it is possible to use a filter *before* the mixer in a receiver to eliminate unwanted signals at the image frequency for the downconversion case. Alternatively, a single-sideband mixer for the transmit path, or an image-reject mixer for the receive path can be used.

An example of a single-sideband upconversion mixer is shown in Figure 8.29. It consists of two basic mixer circuits, two 90° phase shifters, and a summing stage.

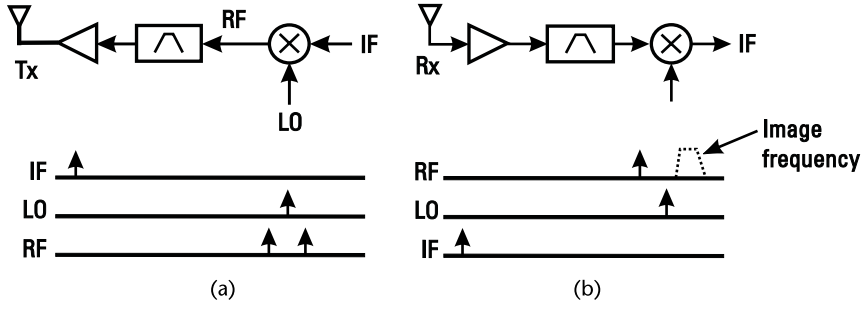

Figure 8.28 (a) Sidebands in upconversion and (b) image in downconversion.

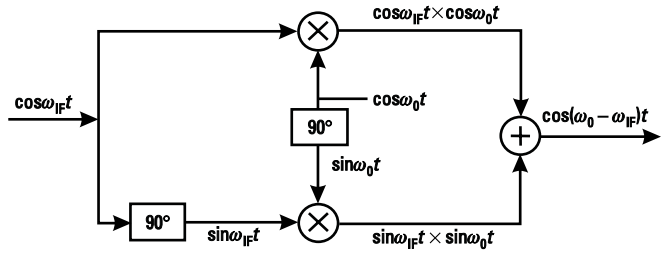

Figure 8.29 A single-sideband mixer.

As can be shown, the use of the phase shifters and mixers will cause one sideband to add in phase and the other to add in antiphase, leaving only the desired sideband at the output. Which sideband is rejected depends on the placement of the phase shifts or the polarity of the summing block. By moving the phase shift from the input to the output, as shown in Figure 8.30, an image-reject mixer is formed. In this circuit, at the output, the RF signal adds in phase while the image adds in antiphase.

8.11.1 Alternative Single-Sideband Mixers

The image-reject configuration in Figure 8.30 is also known as the *Hartley architecture*.

Another possible implementation of an image-reject receiver is known as the *Weaver architecture* shown in Figure 8.31. In this case, the phase shifter after the mixer in Figure 8.30 is replaced by another set of mixers to perform an equivalent operation. The advantage is that all phase shifting takes place only in the LO path and there are no phase shifters in the signal path. As a result, this architecture is less sensitive to amplitude mismatch in the phase-shifting networks and so image rejection is improved. The disadvantage is the additional mixers required, but if the receiver has a two-stage downconversion architecture, then these mixers are already present so there is no penalty.

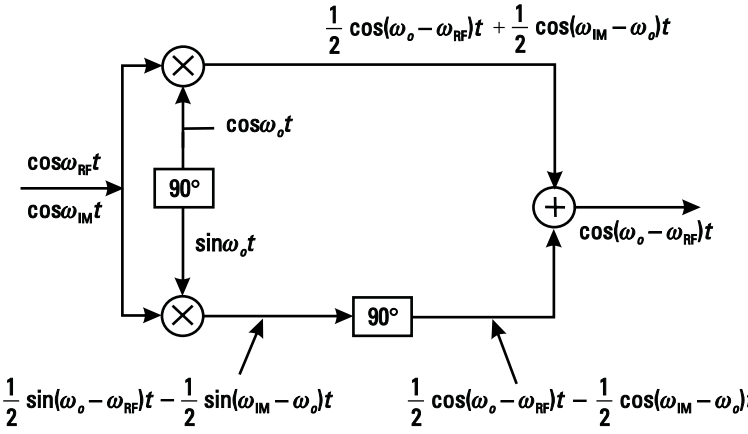

Figure 8.30 An image-reject mixer.

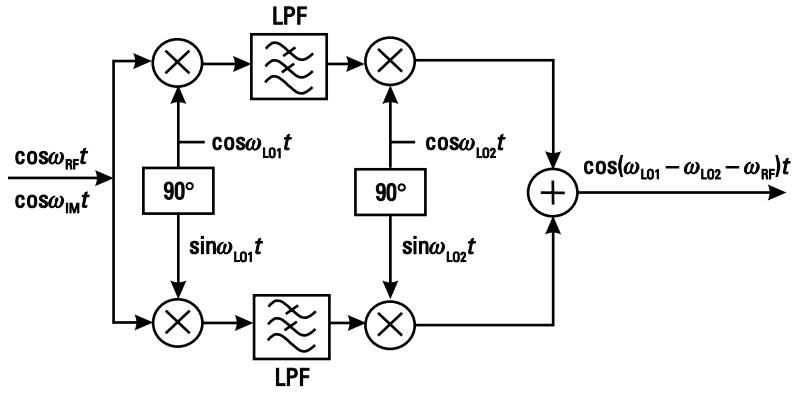

Figure 8.31 Weaver image-reject mixer.

8.11.2 Generating 90° Phase Shift

Several circuits can be used to generate the phase shifts as required for single-sideband or image-reject mixers. Some of the simplest are the RC circuits shown in Figure 8.32. The transfer functions for the two networks are simply

$$\frac{v_{o1}}{v_1} = \frac{sCR}{1+sCR} = \frac{j\omega/\omega_o}{1+j\omega/\omega_o}$$
$$\frac{v_{o2}}{v_2} = \frac{1}{1+sCR} = \frac{1}{1+j\omega/\omega_o} \tag{8.54}$$

where $\omega_o = 1/CR$.

It can be seen that at the center frequency, where $\omega = \omega_o$, the output of the highpass filter is at $v_{o1}/v_1 = 1/\sqrt{2}$ 45° and the output of the lowpass filter is at $v_{o2}/v_2 = 1/\sqrt{2}$ 45°. Thus, if $v_1 = v_2$, then v_{o1} and v_{o2} are 90° out of phase. In

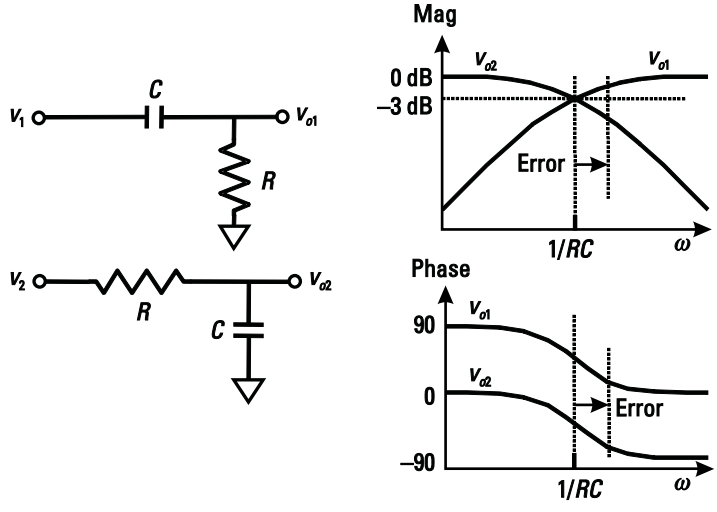

Figure 8.32 RC networks to produce phase shift.

a real circuit, the amplitude or phase may be shifted from their ideal value. Such mismatch between the amplitude or phase can come from a variety of sources. For example, R and C can be poorly matched, and the time constant could be off by a large percentage. As shown in Figure 8.32, such an error will cause an amplitude error, but the phase difference between the two signals will remain at approximately 90°. If the phase-shifted signals are large and fed into the switching quad of a mixer, amplitude mismatch is less important. However, in any configuration requiring a phase shifter in the signal path, such as shown in Figures 8.29 and 8.30, the sideband cancellation or image rejection will be sensitive to amplitude and phase mismatch. Even if the phase shifter is perfect at the center frequency, there will be errors at other frequencies and this will be important in broadband designs.

Example 8.5: Calculation of Amplitude and Phase Error of Phase-Shifting Network

Calculate the amplitude and phase error for a 1% component error.

Solution:
Gains are calculated as

$$\frac{v_{o1}}{v_i} = \frac{j1.01}{1+j1.01} = 0.7106 \quad 44.71°$$

$$\frac{v_{o2}}{v_i} = \frac{1}{1+j1.01} = 0.7036 \qquad 45.29°$$

In this case, the phase difference is still 90°, but the amplitude now differs by about 1%. It will be shown later that such an error will limit the image rejection to about 40 dB.

A differential implementation of a simple phase-shifting circuit is shown in Figure 8.33. In order to function properly, the RC network must not load the out-

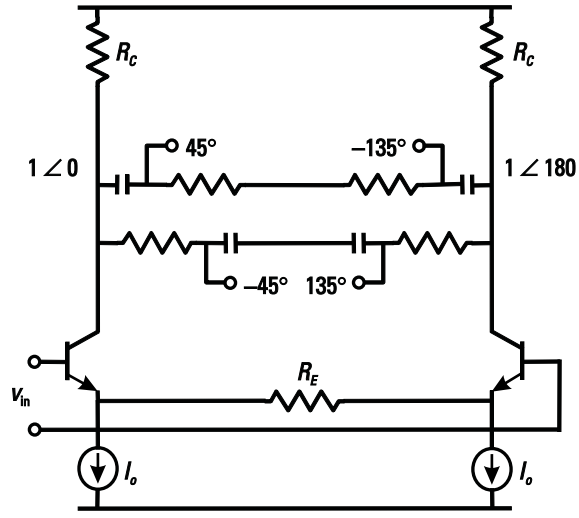

Figure 8.33 Differential circuit to produce phase shift.

put of the differential amplifier. It may also be necessary to buffer the phase shift output.

This circuit is sometimes known as a first-order polyphase filter. The polyphase filter will be discussed in the next section.

Polyphase Filters

A multistage polyphase filter [1] is a circuit technique to improve performance in the presence of component variations and mismatches and over a broader band of frequencies. All polyphase filters are simple variations or extensions of the polyphase filters shown in Figure 8.34. One of the variations is in how the input is

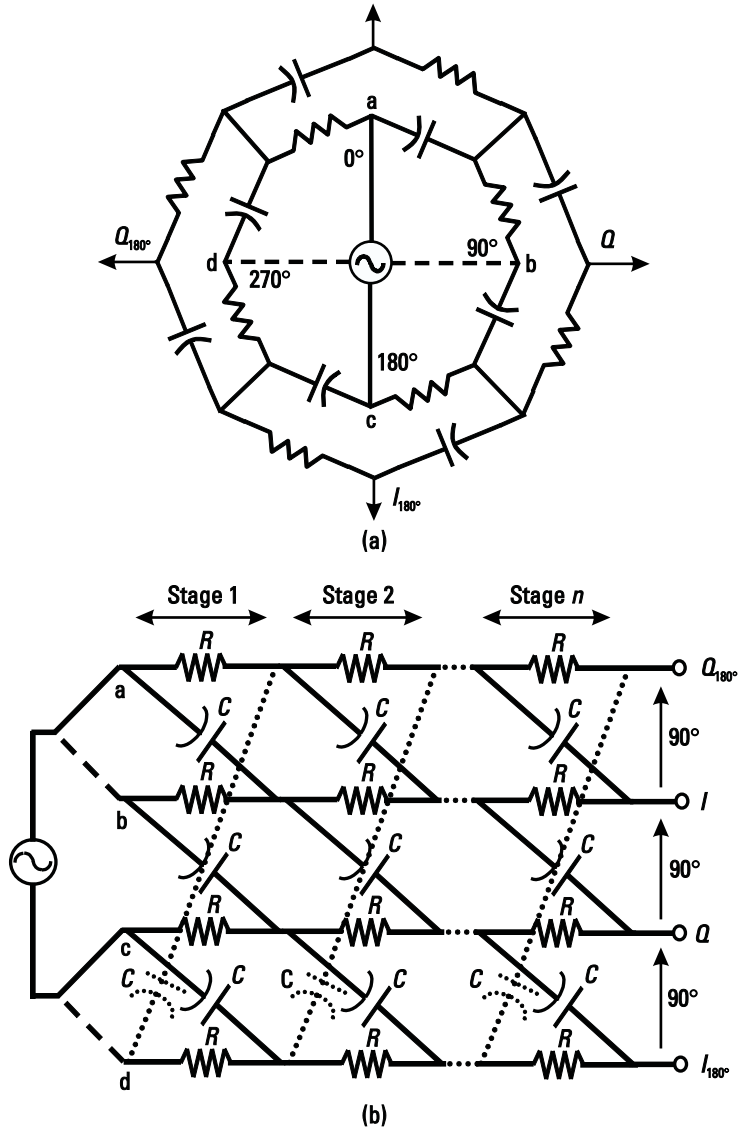

Figure 8.34 Polyphase filters: (a) two-stage and (b) *n*-stage.

driven. The inputs can be driven with four-phases, or simple differential inputs can be applied at nodes a and c. With the simple differential inputs, the other nodes, nodes b and d can be connected to a and c, left open or grounded.

The polyphase filter is designed such that at a particular frequency, (nominally at $\omega = 1/RC$) all outputs are 90° out of phase with each other. The filter also has the property that with each additional stage, phase shifts become more precisely 90° even with a certain amount of tolerance on the parts. Thus, when they are used in an image-reject mixer, if more image rejection is required, then polyphase filters with more stages can be employed. The drawback is that with each additional stage there is an additional loss of about 3 dB through the filter. This puts a practical upper limit on the number of stages that can be used.

8.11.3 Image Rejection with Amplitude and Phase Mismatch

The ideal requirements are that a phase shift of exactly 90° is generated in the signal path and that the LO has perfect quadrature output signals. In a perfect system, there is also no gain mismatch in the signal paths. In a real circuit implementation, there will be imperfections as shown in Figure 8.35. Therefore, an analysis of how much image rejection can be achieved for a given phase and amplitude mismatch is now performed.

The analysis proceeds as follows:

1. The input signal is mixed with the quadrature LO signal through the I and Q mixers to produce signals V_1 and V_2 after filtering. V_1 and V_2 are given by

$$V_1 = \frac{1}{2}\cos(\omega_{\mathrm{LO}} \quad \omega_{\mathrm{RF}})t + \frac{1}{2}\cos(\omega_{\mathrm{IM}} \quad \omega_{\mathrm{LO}})t \tag{8.55}$$

$$V_2 = \frac{1}{2}\sin[(\omega_{\mathrm{LO}} \quad \omega_{\mathrm{RF}})t + \phi_{\varepsilon 1}] \quad \frac{1}{2}\sin[(\omega_{\mathrm{IM}} \quad \omega_{\mathrm{LO}})t + \phi_{\varepsilon 1}] \tag{8.56}$$

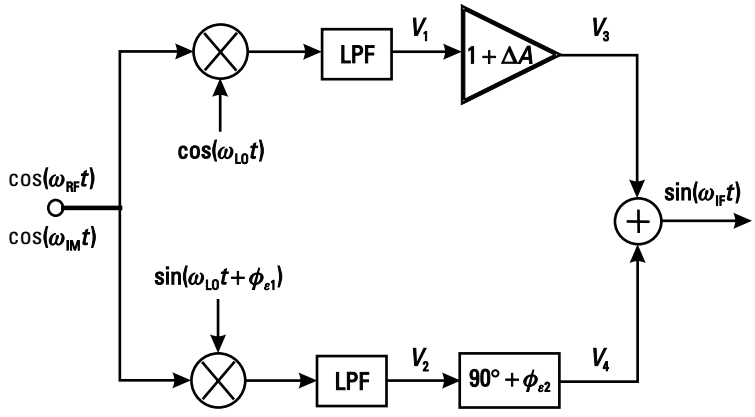

Figure 8.35 Block diagram of an image-reject mixer including phase and gain errors.

2. Now V_1 experiences an amplitude error relative to V_2 and V_2 experiences a phase shift that is not exactly 90° to give V_3 and V_4, respectively.

$$V_3 = \frac{1}{2}(1 + A)\cos(\omega_{LO} - \omega_{RF})t + \frac{1}{2}(1 + A)\cos(\omega_{IM} - \omega_{LO})t \tag{8.57}$$

$$V_4 = \frac{1}{2}\cos[(\omega_{LO} - \omega_{RF})t + \phi_{\varepsilon 1} + \phi_{\varepsilon 2}] - \frac{1}{2}\cos[(\omega_{IM} - \omega_{LO})t - \phi_{\varepsilon 1} + \phi_{\varepsilon 2}] \tag{8.58}$$

3. Now V_3 and V_4 are added together. The component of the output due to the RF signal is denoted V_{RF} and is given by:

$$V_{RF} = \frac{1}{2}(1 + A)\cos(\omega_{IF}t) + \frac{1}{2}\cos(\omega_{IF}t + \phi_{\varepsilon 1} + \phi_{\varepsilon 2}) \tag{8.59}$$

$$V_{RF} = \frac{1}{2}(1 + A)\cos(\omega_{IF}t) + \frac{1}{2}\cos(\omega_{IF}t)\cos(\phi_{\varepsilon 1} + \phi_{\varepsilon 2}) - \frac{1}{2}\sin(\omega_{IF}t)\sin(\phi_{\varepsilon 1} + \phi_{\varepsilon 2}) \tag{8.60}$$

4. The component due to the image is denoted V_{IM} and is given by:

$$V_{IM} = \frac{1}{2}(1 + A)\cos(\omega_{IF}t) - \frac{1}{2}\cos(\omega_{IF}t)\cos(\phi_{\varepsilon 2} - \phi_{\varepsilon 1}) + \frac{1}{2}\sin(\omega_{IF}t)\sin(\phi_{\varepsilon 2} - \phi_{\varepsilon 1}) \tag{8.61}$$

5. Only the ratio of the magnitudes is important. The magnitudes are given by:

$$|V_{RF}|^2 = \frac{1}{4}[\sin(\phi_{\varepsilon 1} + \phi_{\varepsilon 2})]^2 + [(1 + A) + \cos(\phi_{\varepsilon 1} + \phi_{\varepsilon 2})]^2 \tag{8.62}$$

$$|V_{RF}|^2 = \frac{1}{4}[1 + (1 + A)^2 + 2(1 + A)\cos(\phi_{\varepsilon 1} + \phi_{\varepsilon 2})] \tag{8.63}$$

$$|V_{IM}|^2 = \frac{1}{4}(\sin(\phi_{\varepsilon 2} - \phi_{\varepsilon 1}))^2 + [(1 + A) - \cos(\phi_{\varepsilon 2} - \phi_{\varepsilon 1})]^2 \tag{8.64}$$

$$|V_{IM}|^2 = \frac{1}{4}[1 + (1 + A)^2 - 2(1 + A)\cos(\phi_{\varepsilon 2} - \phi_{\varepsilon 1})] \tag{8.65}$$

6. Therefore the image rejection ratio is given by:

$$IRR = 10\log\frac{|V_{RF}|^2}{|V_{IM}|^2} = 10\log\frac{1 + (1 + A)^2 + 2(1 + A)\cos(\phi_{\varepsilon 1} + \phi_{\varepsilon 2})}{1 + (1 + A)^2 - 2(1 + A)\cos(\phi_{\varepsilon 2} - \phi_{\varepsilon 1})} \tag{8.66}$$

If there is no phase imbalance or amplitude mismatch, then this equation approaches infinity, and so ideally, this system will reject the image perfectly and it is only the nonideality of the components that cause finite image rejection. Figure 8.36 shows plots of how much image rejection can be expected for various levels of phase and amplitude mismatch. An amplitude error of about 20% is acceptable

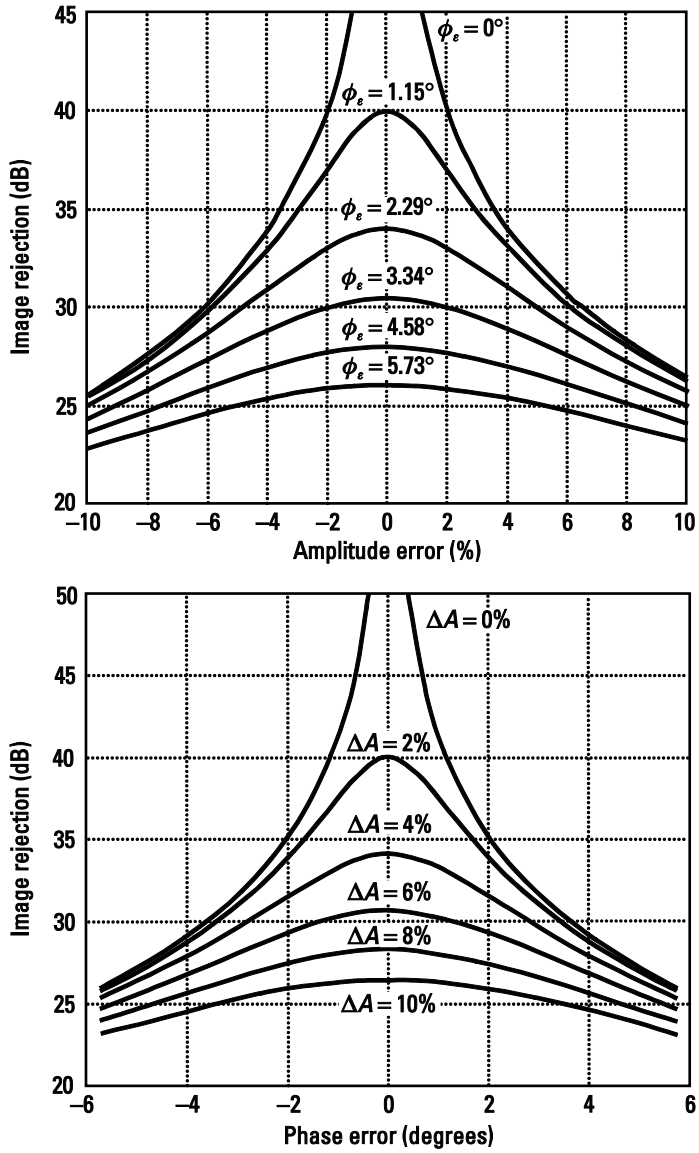

Figure 8.36 Plot of image rejection versus phase and amplitude mismatch.

for 20 dB of image rejection, but more like 2% is required for 40 dB of image rejection. Likewise, phase mismatch must be held to less than 1.2° for 40 dB of image rejection, phase mismatch of less than 11.4° can be tolerated for 20 dB of image rejection.

8.12 Alternative Mixer Designs

In the following section, some variations of mixers will be mentioned briefly, including the Moore mixer, which rejects image noise from a degeneration resistor,

mixers which make use of inductors and transformers, and some other low voltage mixers.

8.12.1 The Moore Mixer

In a receiver, the noise produced by the mixer is sometimes very important. If the mixer is to use resistive degeneration and it is to have its phase shifts in the IF and LO paths, then there is a way to interleave the mixers, as shown in Figure 8.37, such that the noise produced by the degeneration resistors R_E is also image-rejected. Here, the noise due to these resistors is fed into both paths of the mixer rather than just one; thus it gets image-rejected, and its effect is reduced by 3 dB. Since noise due to degeneration resistors is often very important, this can have a beneficial effect on the noise figure of the mixer.

8.12.2 Mixers with Transformer Input

Figure 8.38 shows a mixer with a transformer-coupled input and output [2]. Such a mixer has the potential to be highly linear, since a transformer is used in place of the input transistors. In addition, this mixer can operate from a low power supply voltage, since the number of stacked transistors is reduced compared to that of a conventional mixer. We note that for a downconversion mixer, the input transformer could be on-chip, but for low IF, the output transformer would have to be off-chip.

8.12.3 Mixer with Simultaneous Noise and Power Match

Figure 8.39 shows a bipolar mixer with inductor degeneration and inductor input achieving simultaneous noise and power matching similar to that of a typical LNA [3]. This is also equally useful for CMOS mixers.

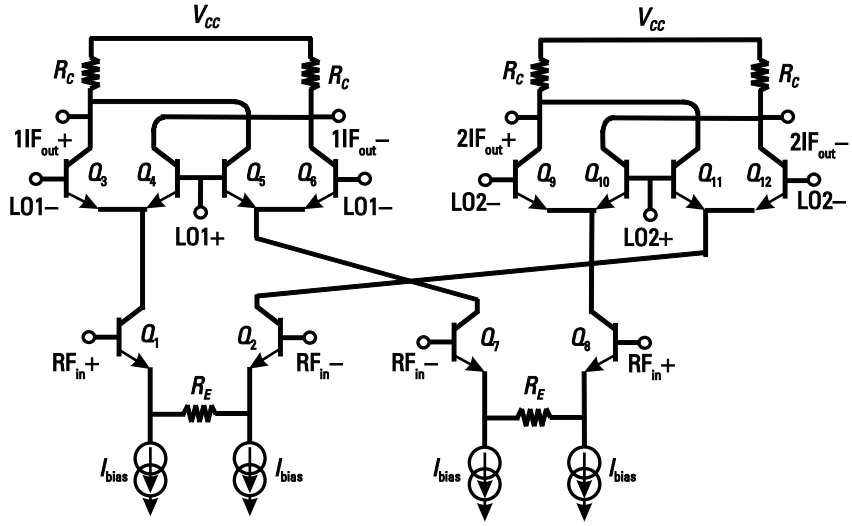

Figure 8.37 The Moore mixer.

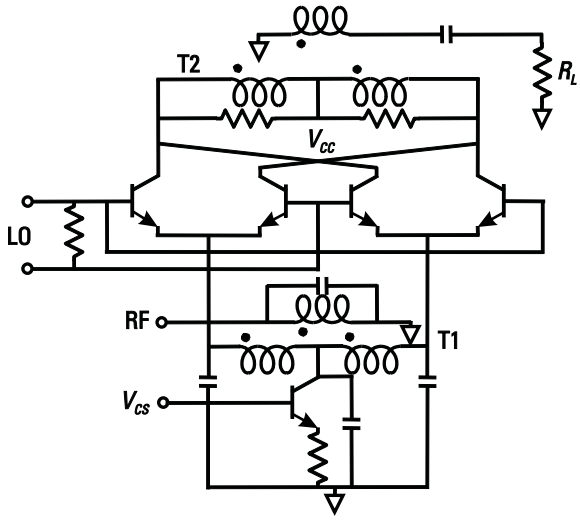

Figure 8.38 Mixer with transformer input.

To achieve matching, the same conditions as for an LNA are required, starting with:

$$L_E = \frac{Z_0}{2\pi f_T} \tag{8.67}$$

The resulting linearity is approximately given by:

$$\text{IIP3} \quad \frac{\omega g_m Z_0}{2\pi f_T} \tag{8.68}$$

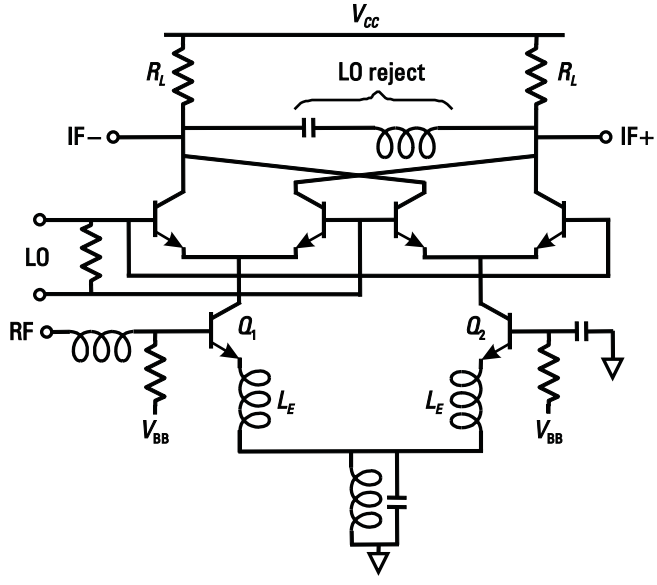

Figure 8.39 Mixer with simultaneous noise and power match.

Noise matching is achieved by sizing L_E, selecting transistor size, and operating the RF transistors at the current required for minimum noise figure. The quad switching transistors are sized for maximum f_T, which typically means they will end up being about 5 to 10 times smaller than the RF transistors.

8.12.4 Mixers with Coupling Capacitors

If headroom is a problem, but due to space or bandwidth constraints it is not possible to use transformers or inductors, then the circuit shown in Figure 8.40, or its CMOS equivalent, may be one alternative. In this figure, the differential amplifier is coupled into the switching quad through the capacitors C_{cc}. Resistors R_{cc} provide a high impedance, so that most of the small-signal current will flow through the capacitors and up into the quad. Usually it is sufficient for R_{cc} to be about ten times the impedance of the series combination C_{cc} and input impedance of the switching quad transistors. Also, the current is steered away from the load resistors using two PMOS transistors. Thus, R_c can be large to give good gain without using up as much headroom as would otherwise be required. A current source I_{BA} can also be included for bias adjustment if needed.

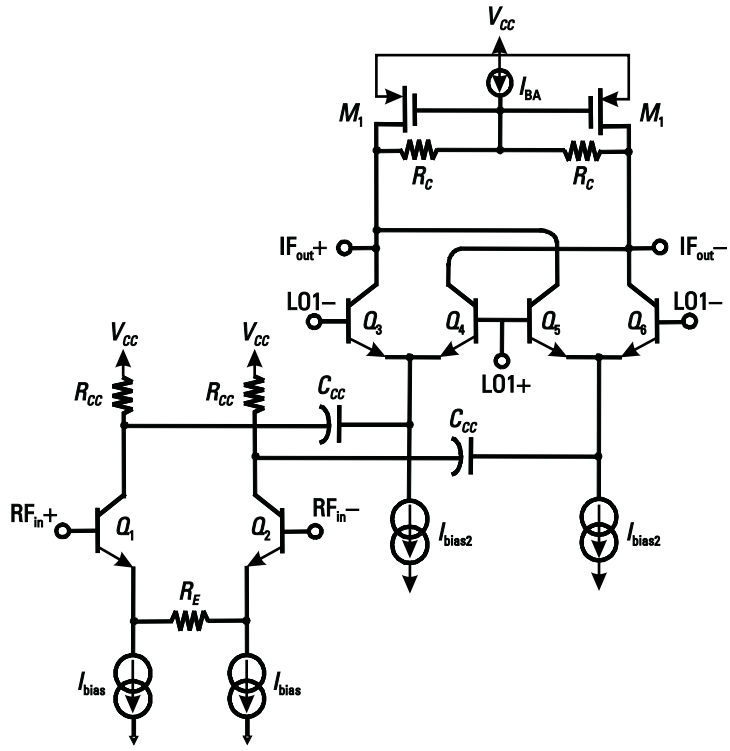

Figure 8.40 Mixer with folded switching stage and current steering PMOS transistors for high gain at low supply voltage.

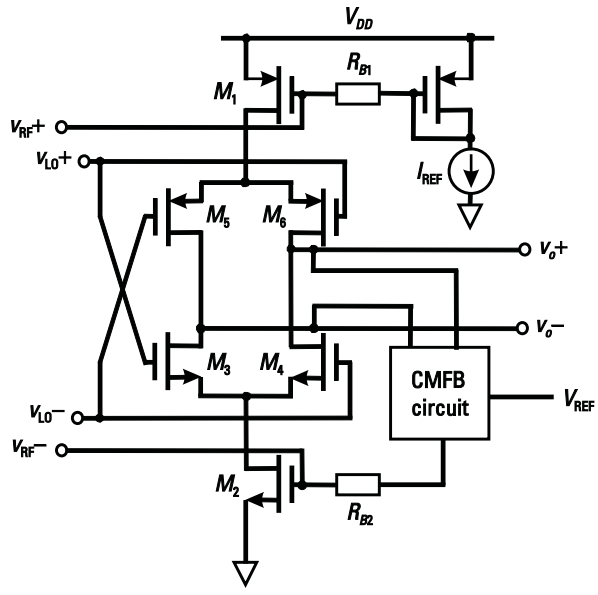

Figure 8.41 CMOS mixer with NMOS and PMOS differential pairs.

8.12.5 CMOS Mixer with Current Reuse

An opportunity with CMOS is to replace an NMOS differential pair with a PMOS differential pair in the RF input and the quad network, which allows them to be stacked and the current to be reused as shown in Figure 8.41 [4]. In such a case, the output is potentially a high-gain node, so some form of common mode feedback is required for this circuit.

8.12.6 Integrated Passive Mixer

The mixers discussed so far depend on a cross-coupled set of four transistors to switch the current coming from the input stage. An integrated passive mixer instead uses a similar set of four transistors to switch the voltage as shown in Figure 8.42. In spite of making use of transistors, this is called a passive mixer because there is

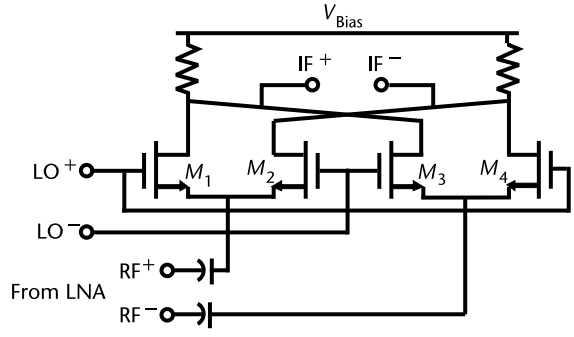

Figure 8.42 Integrated passive mixer.

no dc current flowing through the transistors; in fact, the RF inputs are typically ac coupled. Because there is no dc current, it is possible to operate this mixer at very low power and with low 1/*f* noise. The main disadvantage is that this mixer does not have gain. Since the switching transistors alternately connect the RF inputs to the IF output or to the negative of the IF output, this is equivalent to alternately multiplying the input by +1 or 1. Thus the conversion gain of the mixer is given as

$$\frac{V_{\mathrm{IF}}}{V_{\mathrm{RF}}} = \frac{2}{\pi} \tag{8.69}$$

As a result, the conversion loss of the passive mixer is about 3.9 dB. With nonideal switching the gain expression is much more complex [5]; however, the above is a good starting point for design. Similar to the previous mixer designs, transistors are sized to provide good quality switches and input LO drive is chosen to provide good switching to maximize linearity and to minimize noise. The bias voltage V_{Bias} can be used to bias the following stage, or it can be connected to ground if the following stage is ac coupled.

8.12.7 Subsampling Mixer

The principle of a subsampling mixer is shown in Figure 8.43. Frequencies have been added to illustrate the principle. The input is a sine wave at 1.6 GHz. The input signal is being sampled by a 750 MHz periodic train of impulses, whose frequency spectrum is also a train of impulses. The result is that the sampled input in the frequency domain now contains replications of the input signal at all sum and difference frequencies between the input signal and multiples of the sampling frequency, that is, at $\pm f_R \pm nf_S$. The fundamental or lowest frequency component at $f_R - 2\,f_S$ or 100 MHz can be seen by joining the tips of the impulse samples. Applying a sample and hold applies lowpass filtering reducing the amplitude of higher frequency components leaving the fundamental at 100 MHz as the dominant component. Thus sampling, typically performed by a track and

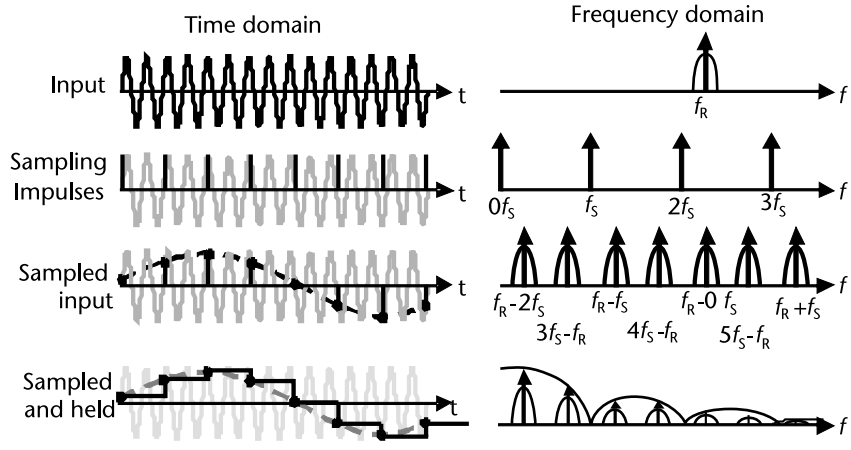

Figure 8.43 Time and frequency domain representation of subsampling with impulses.

hold circuit, samples the carrier at less than Nyquist, thus it is undersampling the carrier. However, the bandwidth of the signal should be less than half the sampling frequency; hence the signal is properly sampled. Because the sampling circuit is operating at lower frequency compared to a conventional mixer, the circuit can be low power. As well, it can be made to have high linearity; however, noise tends to be high because noise at all multiples of the sampling frequency is folded into the baseband.

An example of a sample and hold circuit is shown in Figure 8.44, showing first the track mode then the hold mode [5]. In track mode, M_1 through M_5 are on, and the RF input is transferred to the capacitors with the other side held at the common-mode voltage V_{CM}. The input sampling circuit is designed to be high speed to follow the input signal, but the opamp does not need to track during this phase, so its design can be simplified. During the hold mode, the input and the common-mode voltages are disconnected while M_6 and M_7 are on connecting the opamp to the circuit. During this phase, the opamp can respond slowly at the clock rate, not at the RF input rate, and so it is possible to design the opamp with very low power consumption.

Example 8.6: High-Linearity Mixer

Design a bipolar mixer to downconvert a 2-GHz RF signal to a 50-MHz IF. Use a low-side-injected LO at 1.95 GHz. Design the mixer to have an IIP3 of 8 dBm and 15 dB of voltage gain. The mixer must operate from a 3.3-V supply and draw no more than 12 mA of current. Determine the noise figure of the design as well. Determine what aspects of the design dominate the noise figure. Do not use any inductors in the design and match the input to 100 differentially.

Solution:

Since inductors are not allowed in the design, the linearity must be achieved with resistor degeneration. Since current sources require at least 0.7V and the differential pair and quad will both require 1V, this would leave only 0.6V for the load

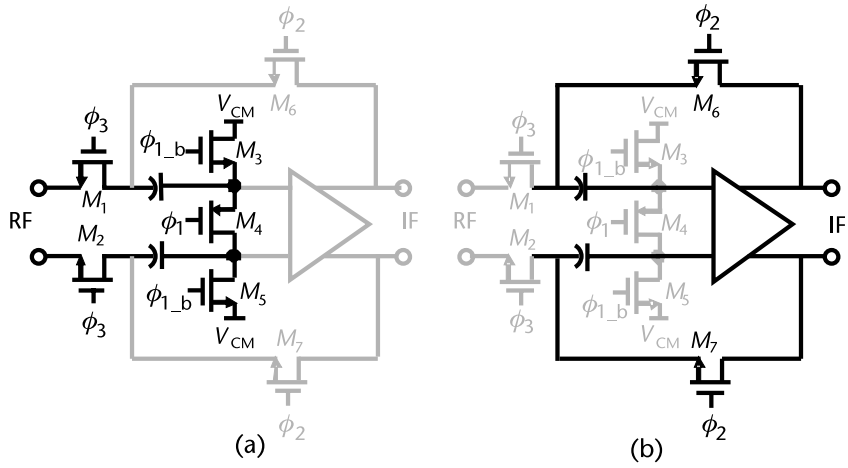

Figure 8.44 Sample and hold circuit (a) in track mode and (b) in hold mode.

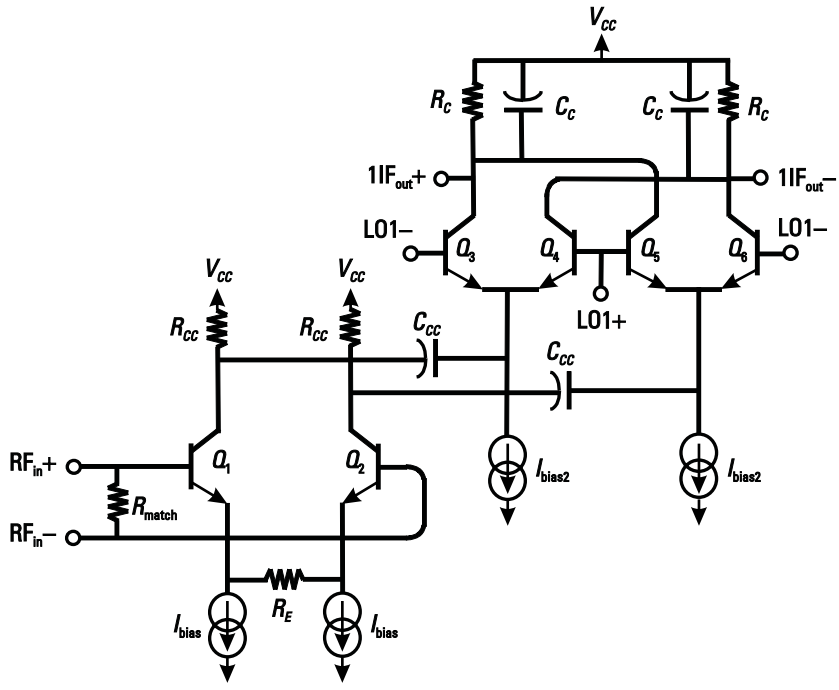

Figure 8.45 Mixer with folded switching stage and resistive input matching.

resistors. A design that stacks the entire circuit is unlikely, therefore, to fit into the 3.3-V supply requirement; thus it will have to be folded. Also, since we are using resistive degeneration, we can probably match the input with a simple resistor. Thus, the mixer topology shown in Figure 8.45 will probably be adequate for this design.

We can now begin sizing components and determining bias currents. First, we are told that we can use 12 mA in this design. There are two sources of nonlinearity of concern, one is the exponential nonlinearity of the differential pair, and the other is the exponential nonlinearity of the quad. We note that the quad nonlinearity for the folded cascode configuration is slightly more complex than the standard configuration because the current applied to it is no longer exactly equal to the current from the RF stage. We can start by assuming that each nonlinearity contributes equally to the linearity of the circuit (assuming that the circuit now has enough headroom that the output does not clip or saturate the quad). Then the input should be designed for 11 dBm IIP3 rather than 8 dBm.

The quad will be more linear as more current is passed through it because of the reduction in the emitter resistance r_e of the four transistors and the resulting reduction of the voltage swing at the emitters. Thus, as a first cut we will split the available current equally between both stages, allotting 6 mA to the driver, and 6 mA to the quad. We can now start to size the degeneration resistor R_E. An IIP3 of 11 dBm at 100 corresponds to a signal swing of 1.12 V_{rms} at the input of the mixer, or equivalently, 0.561 V_{rms} per side for the differential circuit. Using (8.49) and noting that peak differential voltage is 1.58V, R_E can be determined to be:

$$R_{EE} = \frac{v_{\text{IP3}}}{3I_{EE}} \quad 2r_e = \frac{1.58}{3 \quad 6\text{ mA}} \quad 16.6 = 71.2$$

Since this formula is an approximation and we have already included the effect of another nonlinearity, we will choose $R_E = 70$ for this design.

Next, we can find the load resistor noting that we want 15 dB of voltage gain or 5.6 volts-per-volt (V/V). Using (8.28) (omitting output matching):

$$R_C = \frac{\pi}{2} A_v \quad r_e + \frac{R_E}{2} \div = \frac{\pi}{2} 5.6(8.3 \quad + 35 \quad) = 380$$

Note that there will be losses due to the r_o of the quad transistors and some loss of current between the stages, but we will start with an R_C value of 400 . We also include capacitors C_C in parallel with the resistors R_C to filter out high-frequency signals coming out of the mixer. We will choose the filter to have a corner frequency of 100 MHz; therefore, the capacitors should be sized to be

$$C_C = \frac{1}{2\pi f_c R_C} = \frac{1}{2\pi(100\text{ MHz})(400 \quad)} = 4\text{ pF}$$

Now the coupling network needs to be designed. The current sources I_{bias} will need about 0.7V across them to work properly, and the differential pair should have roughly 1.5V to avoid nonlinearity. This leaves the resistors R_{CC} with about 0.8V, thus a value between 200 and 400 would be appropriate for these resistors. We choose 300 .

The quad transistors will each have an r_e of 16.7 . This is less than one-tenth of the value of R_{CC}, and if they are placed in series with a 3-pF capacitor, they still have an impedance with a magnitude of 31.5 or about one-tenth that of R_{CC}. Thus, little current will be lost through the resistors R_{CC}.

The quad transistors themselves were sized so that when operated at 1.5 mA each they were at the current for peak f_T. For minimum noise, the differential pair transistors were sized somewhat larger than the quad transistors. However, since the noise will be dominated by R_E, exact sizing for minimum noise was not critical.

The circuit also needs to be matched. Since inductors have not been allowed, we do this in a crude manner by placing a 100 resistor across the input.

Next, we can estimate the noise figure of this design. The biggest noise contributors will be R_E, R_{Match}, and the source resistance. The noise spectral density produced by both the matching resistor and the source $v_{n(\text{source})}$ will be

$$v_{n(\text{source})} = \sqrt{4kTR_{\text{Match}}} = 1.29\frac{\text{nV}}{\sqrt{\text{Hz}}}$$

These two noise sources are voltage divided at the input by the source and matching resistors. They will also see the same gain to the output. Thus, the output noise generated by each of these two noise sources $v_{on(\text{source})}$ is:

$$v_{no(\text{source})} = \frac{v_{n(\text{source})}}{2} A_v = \frac{1.29 \frac{\text{nV}}{\sqrt{\text{Hz}}}}{2} \times 5.6 \frac{\text{V}}{\text{V}} = 3.6 \frac{\text{nV}}{\sqrt{\text{Hz}}}$$

The other noise source of importance is R_E. It produces a current $i_{n(\text{RE})}$ of

$$i_{n(R_E)} = \sqrt{\frac{4kT}{R_E}} = 15.3 \frac{\text{pA}}{\sqrt{\text{Hz}}}$$

Since R_E is significantly larger than r_e this current will mostly all go to the output, producing an output voltage $v_{no(R_E)}$ of

$$v_{no(R_E)} = \frac{2}{\pi} i_{n(R_E)} \times 2R_C = 7.81 \frac{\text{nV}}{\sqrt{\text{Hz}}}$$

Now the total output noise voltage $v_{no(\text{total})}$ (assuming these are the only noise sources in the circuit) is

$$v_{no(\text{total})} = \sqrt{(v_{\text{no}(R_E)})^2 + (v_{\text{no(source)}})^2 + (v_{\text{no\_source}})^2} = 9.32 \frac{\text{nV}}{\sqrt{\text{Hz}}}$$

Thus, the single-sideband noise figure can be calculated by

$$\text{NF} = 20 \times \log \frac{v_{no(\text{total})} \div}{\frac{v_{on(\text{source})}}{\sqrt{2}} \div} = 20 \times \log \frac{9.32}{2.54} \div = 11.3 \text{ dB}$$

Note that in single-sideband noise figure, only the source noise from one sideband is considered; thus we divided by $\sqrt{2}$.

Now the circuit is simulated. The results are summarized in Table 8.3. The voltage gain was simulated to be 13.6 dB, which is 1.4 dB lower than what was calculated. The main source of error in is this calculation is ignoring current lost into R_{cc}. Since the impedance of R_{cc} is about 10 times that of the path leading into the quad, it draws one-tenth of the total current causing a 1-dB loss in gain. Thus in a second iteration R_c would have to be raised to a higher value. The noise figure was also simulated and found to be 12.9 dB. This is close to what was calculated. Most noise came either from R_E, or from both the source and the input matching resistor. A more refined calculation taking more noise sources into account would have made the calculation agree much closer with simulation. To determine the

Table 8.3 Results of the Simulation of the Mixer Circuit

Parameter	*Value*
Gain	13.6 dB
(SSB) NF	12.9 dB
IIP3	8.1 dBm
Voltage	3.3V
Current	12 mA
LO frequency	1.95 GHz
RF	2 GHz
IF	50 MHz

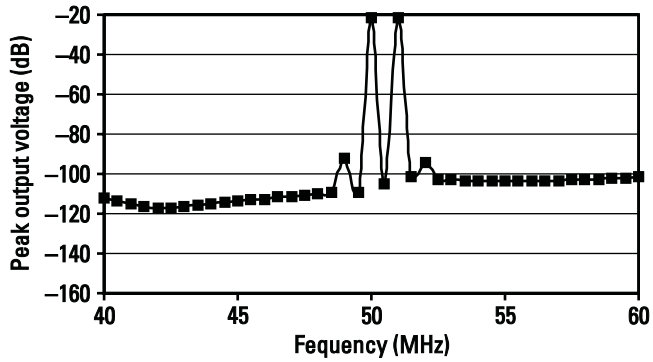

Figure 8.46 FFT of a transient simulation with two input tones used to find the IIP3.

IIP3, the LO was set to be 400 mV_{pp} at 1.95 GHz, and two RF signals were injected at 2.0 and 2.001 GHz. The *fast Fourier transform* (FFT) of the output voltage is plotted in Figure 8.46. From this figure, using a method identical to that used in the broadband LNA example in Section 7.9, it can be found that the IIP3 is 8.1 dBm. Thus, simulations are in good agreement with the calculations.

Example 8.7: Image-Reject Mixer

Take the balanced mixer cell designed in the last example and use it to construct an image-reject mixer as shown in Figure 8.30. Place a simple lowpass-highpass phase shifter in the LO path. Place the second phase shifter in the IF path and make this one a second-order polyphase filter. Compare the design to one using only a first-order polyphase filter. Design the mixer so that it is able to drive 100 output impedance. Explore the achievable image rejection over process tolerances of 20%.

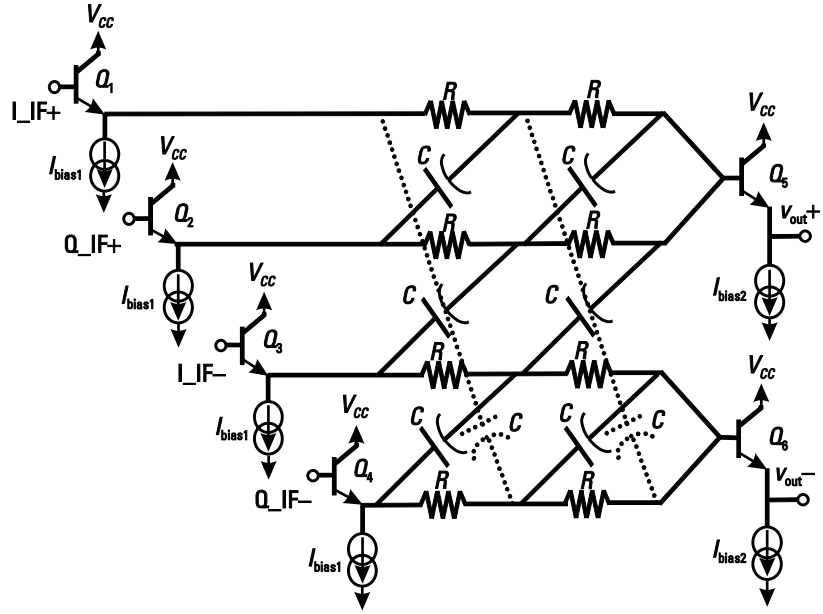

Figure 8.47 IF stage of an image-reject mixer.

Solution:

For the LO path there is little additional design work to be done. For this example, we will ignore the square wave buffers that would normally be used to guarantee the mixer is driven properly. We add a simple phase shifting filter to the circuit like the one shown in Figure 8.33 to provide quadrature LO signals. Since this filter must be centered at 1.95 GHz we choose $R = 300$ fairly arbitrarily and this makes the capacitors 272 fF. Both of these are easily implemented in most technologies.

Next we must design the IF stage that will follow the mixers using the polyphase filter to achieve the second phase shifter. In order to prevent loading of the mixers by the polyphase filter, we need buffers at the input and we will need buffers at the output to drive the 100 load impedance. A polyphase filter with buffers is shown in Figure 8.47. Note that this circuit implements both IF paths as well as the summer shown in Figure 8.30. The polyphase filter components must be sized so that the impedance is large to minimize buffer current. However, if the impedance is made too large, then it will form a voltage divider with the output stage resulting in a loss of gain. Thus, we choose through trial and error a resistance of $R = 2$ k and this will make the capacitors $C = 1.6$ pF (centered at 50 MHz in this case).

The mixer of the previous example had a IIP3 of 8.1 dBm. As there are now two mixers we can expect this system to have an IIP3 of more like 5 dBm. This means that it will have a 1-dB compression point of –5 dBm. At this power level, the input will have a peak voltage swing of 250 mV. With a gain of 13.6 dB or 4.8 V/V, this means that the buffers will have to swing 1.2-V peak. If we assume that they drive a series combination of 2 k and 1.6 pF, then the total impedance will be about 2.8 k . This means that the transistor needs to accommodate an ac current of 429 μA. Thus, a bias current of 750 μA for this stage should be safe.

If we assume now that the polyphase filter has a loss of 3 dB per stage, then the voltage gain from input to output will drop to 7.6 dB or 2.4 V/V. Thus, the peak output voltage will be 600 mV. Into 100 , this will be a current of 6 mA. This large value demonstrates how hard it is to drive low impedances with high linearity systems. We will start with a current of 5 mA in each transistor and refine this number as needed.

The circuit was then simulated. The basic circuit parameters are shown in Table 8.4. The gain and IIP3 has dropped as expected. The noise figure has also risen due

Table 8.4 Results of the Simulation of the Image-Reject Mixer Circuit

Parameter	*Value*
Gain	7.4 dB
NF	16.3 dB
IIP3	6.9 dBm
Voltage	3.3V
Current	37 mA
Image rejection	69 dB
RF frequency	2 GHz
LO frequency	1.95 GHz
IF frequency	50 MHz
Image frequency	1.9 GHz

Table 8.5 Image Rejection for LO Phase Shifter

Tolerance Level for Resistance and Capacitance	*Image Rejection*
±20%	>20 dB
±10%	>27.4 dB
Nominal	69 dB

to reduced gain, but not too much, as now the noise due to the input has been image rejected as well.

The components in the filters were then adjusted to show the effect of circuit tolerance on the image rejection. Table 8.5 shows how the LO phase shifter affects image rejection. Note that this port is very insensitive to amplitude changes, which is why the highpass-lowpass filter was chosen for the 90° phase shift. It still provides image rejection of 20 dB even at 20% tolerance in the values.

Table 8.6 shows that the polyphase filter with two stages also does an excellent job at keeping the image suppressed. Thus this was a good choice for the IF filter. If this filter is reduced to a first order as shown in Table 8.7 then the image rejection suffers greatly. Thus, a second-order filter is required in this case.

Example 8.8: Image-Reject Mixer with Improved Gain

The gain of the image-reject mixer has been reduced by 6 dB due to the presence of the IF polyphase filter. Modify it to get the 6 dB of gain back.

Solution:

With the current flowing in the quad stage of the mixer, it would be impossible, due to headroom constraints, to raise the resistance, so we must now employ the PMOS current steering technique shown in Figure 8.40. The PMOS will now make up the capacitor that was placed in the load to remove high frequency feed through of RF and LO signals. The PMOS must be made large to ensure that they are not noisy and that they have a low saturation voltage. A device with a length of 2 μm and a width of 800 μm was chosen through simulation. No current source was needed in this case, as the voltage levels seemed to be fine without it.

The resistors were then doubled to 800 to restore the gain of the circuit. As well, the buffers are all doubled in current because they will now have to handle signals that are twice as large. The results of this new image-reject mixer are shown in Table 8.8.

Table 8.6 Image Rejection for IF Phase Shifter

Tolerance Level for Resistance and Capacitance	*Image Rejection*
±20%	>25 dB
±10%	>36.7 dB
Nominal	69 dB

Table 8.7 Image Rejection for IF Phase Shifter (First Order)

Tolerance Level for Resistance and Capacitance	*Image Rejection*
±20%	>12.5 dB
±10%	>18.2 dB
±5%	>23.2 dB
Nominal	35 dB

Table 8.8 Results of the Image-Reject Mixer with PMOS Current Steering Transistors

Parameter	*Value*
Gain	13.6 dB
NF	13.5 dB
IIP3	5.7 dBm
Voltage	3.3V
Current	50 mA
Image rejection	69 dB

Note that the NF has dropped due to the increased gain. The linearity has been degraded slightly due to the addition of the additional nonlinearity of the output resistance of the PMOS transistors.

One more improvement can be made to this circuit. The mixer can be put into a Moore configuration to reduce the effect of R_E on the noise figure. When this was done, the noise due to R_E reduced to about half its previous value, but because it was responsible for only a few percent of the total noise, the new noise figure was lowered by only 0.5 to 13.0 dB. This is not a dramatic improvement, but as it comes at no additional cost, it is worthwhile. If the gain of this mixer were increased further then the importance of R_E on the noise figure would increase and a greater improvement would be seen.

References

[1] Gingell, M. J., "Single Sideband Modulation Using Sequence Asymmetric Polyphase Networks," *Electrical Communications*, Vol. 48, 1973, pp. 21–25.

[2] Long, J. R., "A Low-Voltage 5.1-5.8GHz Image-Reject Downconverter RFIC," *IEEE J. Solid-State Circuits*, Vol. 35, September 2000, pp. 1320–1328.

[3] Voinigescu, S. P., and M. C. Maliepaard, "5.8GHz and 12.6GHz Si Bipolar MMICs," *Proc. ISSCC,* 1997, pp. 372, 373.

[4] Karanicolas, A. N., "A 2.7-V 900-MHz CMOS LNA and Mixer," *IEEE J. Solid-State Circuits*, Vol. 31, December 1996, pp. 1939–1944.

[5] Lee, T. H., *The Design of CMOS Radio Frequency Integrated Circuits,* 2nd ed., Cambridge, U.K.: Cambridge University Press, 2003.

Selected Bibliography

Larson, L. E., (ed.), *RF and Microwave Circuit Design for Wireless Communications*, 2nd ed., Norwood, MA: Artech House, 1997.

Maas, S. A., *Microwave Mixers*, 2nd ed., Norwood, MA: Artech House, 1993.

Rudell, J. C., et al., "A 1.9-GHz Wide-Band IF Double Conversion CMOS Receiver for Cordless Telephone Applications," *IEEE J. Solid-State Circuits*, Vol. 32, December 1997, pp. 2071–2088.

CHAPTER 9

Voltage Controlled Oscillators

9.1 Introduction

An oscillator is a circuit that generates a periodic waveform whether it be sinusoidal, square, triangular as shown in Figure 9.1, or, more likely, some distorted combination of all three. Oscillators are used in a number of applications where a reference tone is required. For instance, they can be used as the clock for digital circuits or as the source of the LO signal in transmitters. In receivers, oscillator waveforms are used as the reference frequency to mix down the received RF to an IF or to baseband. In most RF applications, sinusoidal references with a high degree of spectral purity (low phase noise) are required. Thus, this chapter will focus on LC-based oscillators, as they are the most prominent form of oscillator used in RF applications.

Perhaps the most important characteristic of an oscillator is its phase noise. In other words, we desire accurate periodicity with all signal power concentrated in one discrete oscillator frequency and possibly at multiples of the oscillator frequency. As well, since oscillators are designed to operate at particular frequencies of interest, long-term stability is of concern, especially in products that are expected to function for many years. Thus, we would like to have minimum drift of oscillation frequency due to such things as aging or power supply variations. In addition, oscillators must produce sufficient output voltage amplitude for the intended application. For instance, if the oscillator is used to drive the LO switching transistors in a double-balanced mixer cell, then the voltage swing must be large enough to switch the mixer.

In this chapter, we will first look at some general oscillator properties and then examine the resonator as a fundamental building block of the oscillator. Different types of oscillators will then be examined, but most emphasis will be on the Colpitts oscillator and the negative transconductance oscillator. Both single-ended and double-ended designs will be considered. This chapter will also include discussions of the theoretical calculations of the amplitude of oscillation, and the phase noise of the oscillator.

9.2 The LC Resonator

At the core of almost all integrated RF oscillators is an LC resonator that determines the frequency of oscillation and often forms part of the feedback mechanism used to obtain sustained oscillations. Thus, the analysis of an oscillator begins with

Figure 9.1 Example of periodic waveforms.

the analysis of a damped LC resonator such as the parallel resonator shown in Figure 9.2.

Since there are two reactive components, this is a second-order system, which can exhibit oscillatory behavior if the losses are low or if energy is added to the circuit. It is useful to find the system's response to an impulse of current, which in a real system could represent noise. If $i(t) = I_{\text{pulse}}\delta(t)$ is applied to the parallel resonator, the time domain response of the system can be found as

$$v_{\text{out}}(t) \quad \frac{I_{\text{pulse}}}{C} e^{\frac{t}{2RC}} \cos\sqrt{\frac{1}{LC} \quad \frac{1}{4R^2C^2}} \div \times t \div \tag{9.1}$$

From this equation, it is easy to see that this system's response is a sinusoid with exponential decay whose amplitude is inversely proportional to the value of the capacitance of the resonator and whose frequency is given by

$$\omega_{\text{osc}} = \sqrt{\frac{1}{LC} \quad \frac{1}{4R^2C^2}} \tag{9.2}$$

which shows that as $|R|$ decreases, the frequency decreases. However, if $|R| >> \sqrt{L/C}$, as is the case in most RFIC oscillators even during startup, this effect can be ignored. Also note that once steady state has been reached in a real oscillator, R approaches infinity and the oscillating frequency will approach

$$\omega_{\text{osc}} = \sqrt{\frac{1}{LC}} \tag{9.3}$$

The resulting waveform is shown in Figure 9.3. To form an oscillator, however, the effect of damping must be eliminated in order for the waveform to persist.

9.3 Adding Negative Resistance Through Feedback to the Resonator

The resonator is only part of an oscillator. As can be seen from Figure 9.3, in any practical circuit, oscillations will die away unless energy is added in order to sustain the oscillation. A feedback loop can be designed to generate a negative resistance to

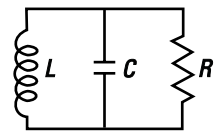

Figure 9.2 Parallel LC resonator.

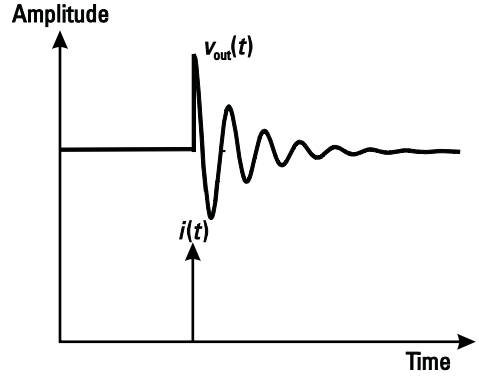

Figure 9.3 Damped LC resonator with current impulse applied.

add energy to the system as shown conceptually in Figure 9.4. If this parallel negative resistance [Figure 9.4(a)] is smaller than the positive parallel resistance in the circuit, then any noise will start an oscillation whose amplitude will grow with time. Similarly, in Figure 9.4(b), if the negative series resistance is larger than the positive resistive losses, then this circuit will also start to oscillate.

The oscillator can be seen as a linear feedback system as shown in Figure 9.5. The oscillator is broken into two parts, which together describe the oscillator and the resonator.

At the input, the resonator is disturbed by an impulse that represents a broadband noise stimulus that starts up the oscillator. The impulse input results in an output that is detected by the amplifier. If the phase shift of the loop is correct and the gain around the loop is such that the pulse that the amplifier produces is equal in magnitude to the original pulse, then the pulse acts to maintain the oscillation amplitude with each cycle. This is a description of the Barkhausen criteria which will now be described mathematically.

The gain of the system in Figure 9.5 is given by

$$\frac{V_{\text{out}}(s)}{V_{\text{in}}(s)} = \frac{H_1(s)}{1 \quad H_1(s)H_2(s)} \tag{9.4}$$

We can see from the equation that if the denominator approaches zero, with finite $H_1(s)$, then the gain approaches infinity and we can get a large output voltage for an infinitesimally small input voltage. This is the condition for oscillation. By solving for this condition, we can determine the frequency of oscillation and the

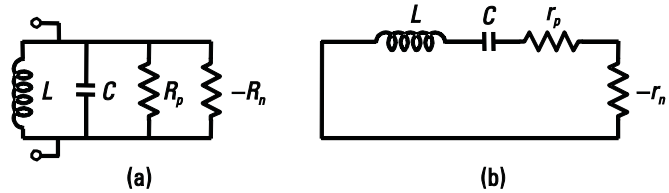

Figure 9.4 The addition of negative resistance to the circuit to overcome losses in (a) a parallel resonator or (b) a series resonator.

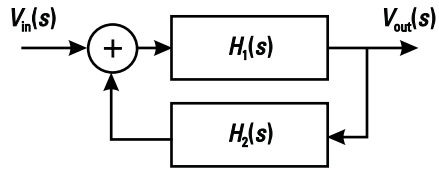

Figure 9.5 Linear model of an oscillator as a feedback control system.

required gain to result in oscillation. Later in the chapter we will see that we can use both an open loop and closed loop transfer function to analyze an oscillator.

More formally, the system poles are defined by the denominator of (9.4). To find the poles of the closed-loop system, one can equate this expression to zero as in

$$1 \quad H_1(s)H_2(s) = 0 \tag{9.5}$$

For sustained oscillation at constant amplitude, the poles must be on the $j\omega$ axis. To achieve this, we replace s with $j\omega$ and set the equation equal to zero.

For the open-loop analysis, rewrite the above expression as

$$H_1(j\omega)H_2(j\omega) = 1 \tag{9.6}$$

Since in general $H_1(j\omega)$ and $H_2(j\omega)$ are complex, this means that

$$\left|H_1(j\omega)\right|\left|H_2(j\omega)\right| = 1 \tag{9.7}$$

and that

$$H_1(j\omega)H_2(j\omega) = 2n\pi \tag{9.8}$$

where n is a positive integer.

These conditions for oscillation are known as the Barkhausen criterion, which states that for sustained oscillation at constant amplitude, the gain around the loop is 1 and the phase around the loop is 0 or some multiple of 2π. We note that H_1H_2 is simply the product of all blocks around the loop and so can be seen as open-loop gain. Also, it can be noted that, in principle, it does not matter where one breaks the loop or which part is thought of as the feedback gain or which part is forward gain. For this reason, we have not specified what circuit components constitute H_1 and H_2 and, in fact, many different possibilities exist.

9.4 Popular Implementations of Feedback to the Resonator

Feedback (which results in negative resistance in the case of many common oscillator topologies) is usually provided in one of three ways as shown in Figure 9.6. (Note that other choices are possible.)

According to the simple theory developed so far, if the overall resistance is negative, then the oscillation amplitude will continue to grow indefinitely. In a practical circuit, this is, of course, not possible. Current limiting, the power supply rails or some nonlinearity in the device eventually limits the amplitude of the oscillation to

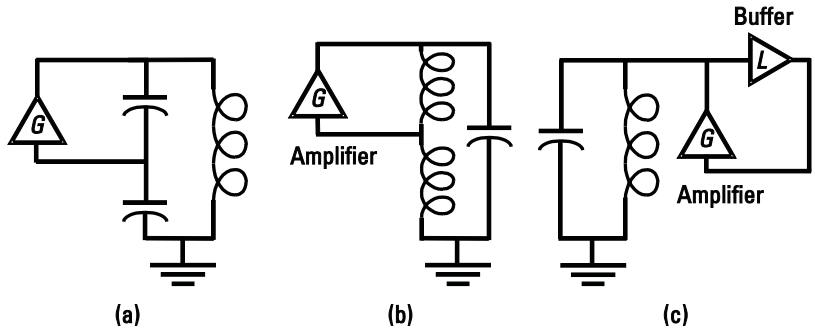

Figure 9.6 Resonators with feedback. Biasing is not shown. (a) Using a tapped capacitor and amplifier to form a feedback loop. This is known as the Colpitts oscillator. (b) Using a tapped inductor and amplifier to form a feedback loop. This is known as the Hartley oscillator. This form is not common in IC implementations. (c) Using two amplifiers (typically two transistors) in a positive feedback configuration. This is known as the G_m oscillator.

some finite value, as shown in Figure 9.7. This reduces the effect of the negative resistance in the circuit until the losses are just canceled, which is equivalent to reducing the loop gain to 1.

9.5 Configuration of the Amplifier (Colpitts or $-G_m$)

The amplifier shown in Figure 9.6 is usually made using only one transistor in RF oscillators. The $-G_m$ oscillator (Figure 9.8) can be thought of as having either a common-drain amplifier made up of M_2, where M_1 forms the feedback, or a common-gate amplifier consisting of M_1 where M_2 forms the feedback. Figure 9.8 may look a little unusual because the $-G_m$ oscillator is usually seen only in a differential form, in which case the two transistors are connected as a differential pair. The circuit is symmetrical when it is made differential (more on this in Section 9.9). However, the Colpitts and Hartley oscillators, each having only one transistor, can be made either common base/gate or common collector/drain. The common emitter/source configuration is usually unsuitable for a completely integrated solution because it requires large capacitors and RF chokes that are not usually available in a typical IC technology. The common emitter/source configuration also suffers from the Miller effect because neither the collector nor the base is grounded. The two favored choices (common base and common collector) are shown in Figure 9.9 as they would appear in the Colpitts oscillator.

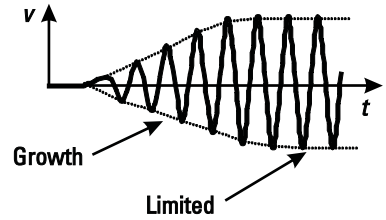

Figure 9.7 Waveform of an LC resonator with losses compensated. The oscillation grows until a practical constraint limits the amplitude.

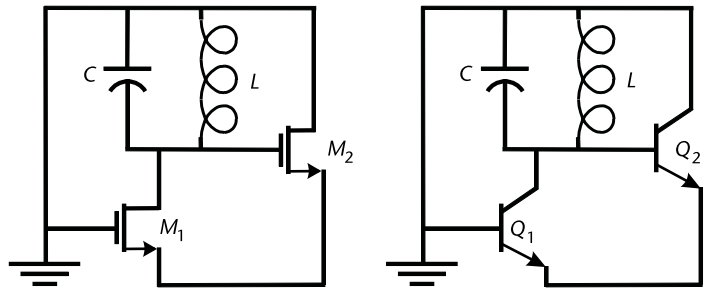

Figure 9.8 G_m oscillator. Biasing not shown.

9.6 Analysis of an Oscillator as a Feedback System

It can be instructive to apply the model of Figure 9.5 to the oscillator circuits discussed above. Expressions for H_1 and H_2 can be found and used in either an open-loop analysis or a closed-loop analysis. For the closed-loop analysis, the system's equations can also be determined, and then the poles of the system can be found. This is the approach we will take first. Later we will demonstrate the open-loop analysis technique. All of these techniques give us two basic pieces of information about the oscillator in question: (1) it allows us to determine the frequency of oscillation, and (2) it tells us the amount of gain required to start the oscillation.

9.6.1 Oscillator Closed-Loop Analysis

In this section, the common-base configuration of the Colpitts oscillator as shown in Figure 9.9 will be considered. The small-signal model of the oscillator is shown in Figure 9.10. Note that the small-signal model would be identical if CMOS transistors were used.

We start by writing down the closed-loop system equations by summing the currents at the collector (node v_c) and at the emitter of the transistor (node v_e). At the collector,

$$v_c \left[\frac{1}{R_p}+\frac{1}{sL}+sC_1\right] - v_e(sC_1+g_m)=0 \tag{9.9}$$

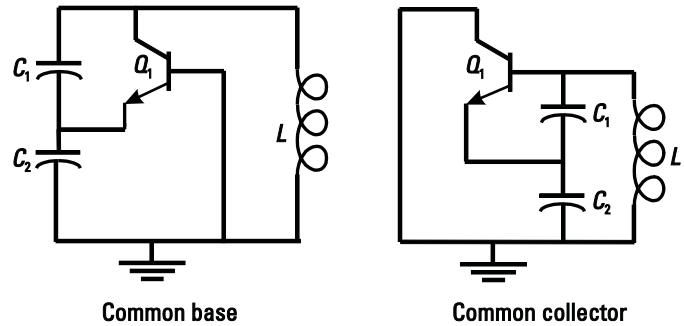

Figure 9.9 Common base and common collector Colpitts oscillators. Biasing not shown.

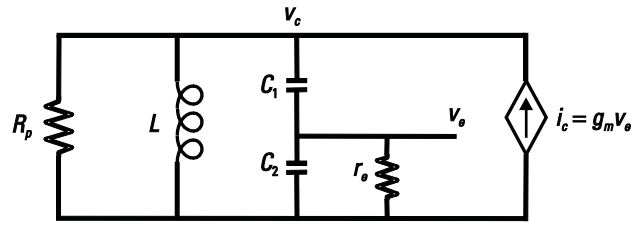

Figure 9.10 Closed-loop oscillator small-signal model.

At the emitter we have:

$$v_e \left(sC_1 + sC_2 + \frac{1}{r_e} \right) - v_c sC_1 = 0 \tag{9.10}$$

This can be solved in several ways; however, we will write it as a matrix expression as:

$$[Y][v] = 0 \tag{9.11}$$

as in the following equation:

$$\begin{matrix} \frac{1}{R_p} + \frac{1}{sL} + sC_1 & sC_1 & g_m \\ sC_1 & sC_1 + sC_2 + \frac{1}{r_e} & \end{matrix} \begin{matrix} v_c \\ v_e \end{matrix} = \begin{matrix} 0 \\ 0 \end{matrix} \tag{9.12}$$

The poles will be formed by the determinant of the matrix. To find the conditions for oscillation, we can set the determinant to zero and solve. The result is:

$$\left(\frac{1}{R_p} + \frac{1}{sL} + sC_1 \right) \left(sC_1 + sC_2 + \frac{1}{r_e} \right) - sC_1(sC_1 + g_m) = 0 \tag{9.13}$$

After multiplying out and collecting like terms, this results in:

$$s^3 LC_1C_2 + s^2 \left(\frac{L(C_1 + C_2)}{R_p} + \frac{LC_1}{r_e} \right) - LC_1 g_m + s \left(\frac{L}{R_p r_e} + C_1 + C_2 \right) + \frac{1}{r_e} = 0 \tag{9.14}$$

Substituting $s = j\omega$ in (9.14) results in even-order terms (s^2 and constant term) being real and odd-order terms (s^3 and s) having a $j\omega$ in them. Thus, when the even-order terms are summed to zero, the result will be an expression for gain. When odd-order terms are summed to zero, to result will be an expression for the frequency. The result for the odd-order terms is:

$$\omega = \sqrt{\frac{C_1 + C_2}{C_1C_2} \frac{1}{L} + \frac{1}{r_e R_p C_1 C_2}} \tag{9.15}$$

The first term can be seen to be ω_o, the resonant frequency of the resonator by itself. The second term can be simplified by noting that ω_o is determined by L

resonating with the series combination of C_1 and C_2 as well as by noting that the Q of an inductor in parallel with a resistor is given by

$$Q_L = \frac{R_p}{\omega L} \tag{9.16}$$

Then:

$$\omega = \sqrt{\omega_o^2 + \frac{\omega_o^2 L}{r_e R_p (C_1 + C_2)}} = \omega_o \sqrt{1 + \frac{\omega L}{R_p} \times \frac{1}{\omega r_e (C_1 + C_2)}} = \omega_o \sqrt{1 + \frac{1}{Q_L \omega / \omega_c}} \tag{9.17}$$

where ω_c is the corner frequency of the highpass filter formed by the capacitive feedback divider.

Thus, if the inductor Q is high or if the operating frequency is well above the feedback corner frequency, then the oscillating frequency is given by ω_o. Otherwise, the frequency is increased by the amount shown. This effect will be revisited in Section 9.6.2 and Example 9.2.

The result for the even-order term in (9.14) is

$$g_m = \frac{\omega(C_1 + C_2)}{Q_L} \tag{9.18}$$

Note that the approximation has been made that $r_e = 1/g_m$. Thus, this equation tells us what value of g_m (and corresponding value of r_e) will result in sustained oscillation at a constant amplitude. For a real oscillator, to overcome any additional losses not properly modeled and to guarantee startup and sustained oscillation at some nonzero amplitude, the g_m would have to be made larger than this value. How much excess g_m is used will affect the amplitude of oscillation. This is discussed further in Section 9.15.

9.6.2 Capacitor Ratios with Colpitts Oscillators

In this section, the role of the capacitive divider as it affects frequency of oscillation and feedback gain will be explored. It will be seen that this capacitor divider is responsible for isolating the loading of r_e on the resonant circuit and produces the frequency shift as mentioned above. The resonator circuit including the capacitive feedback divider is shown in Figure 9.11. Note that the small-signal model would be identical if CMOS transistors were used.

The capacitive feedback divider, made up of C_1, C_2, and r_e, has the transfer function

$$\frac{v_e}{v_c} = \frac{j\omega r_e C_1}{1 + j\omega r_e (C_1 + C_2)} = \frac{C_1}{C_1 + C_2} \div \frac{j\dfrac{\omega}{\omega_c} \div}{1 + j\dfrac{\omega}{\omega_c} \div} \tag{9.19}$$

This is a highpass filter with gain and phase as shown in Figure 9.12.

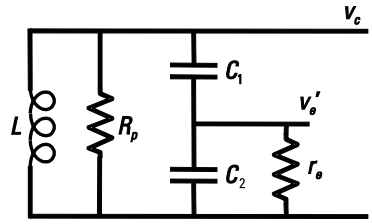

Figure 9.11 Z_{tank} using transformation of capacitive feedback divider.

The passband gain A_o is given by

$$A_o = \frac{C_1}{C_1 + C_2} \tag{9.20}$$

The corner frequency ω_c is given by

$$\omega_c = \frac{1}{r_e(C_1 + C_2)} \tag{9.21}$$

and the phase shift of the feedback network is

$$\phi = \frac{\pi}{2} \quad \tan^{1} \frac{\omega}{\omega_c} \div \tag{9.22}$$

If the frequency of operation is well above the corner frequency ω_c, the gain is given by the capacitor ratio in (9.20) and the phase shift is zero. Under these conditions, the circuit can be simplified as described in the following paragraph. If this frequency condition is not met, there will be implications, which will be discussed later.

This highpass filter also loads the resonator with r_e (the dynamic emitter resistance) of the transistor used in the feedback path. Fortunately, this resistance is transformed to a higher value through the capacitor divider ratio. This impedance transformation effectively prevents this typically low impedance from reducing the Q of the oscillator's LC resonator. The impedance transformation is discussed in Chapter 5 and is given by

$$r_{e,\text{tank}} = \left(1 + \frac{C_2}{C_1} \div\right)^{2} r_e \tag{9.23}$$

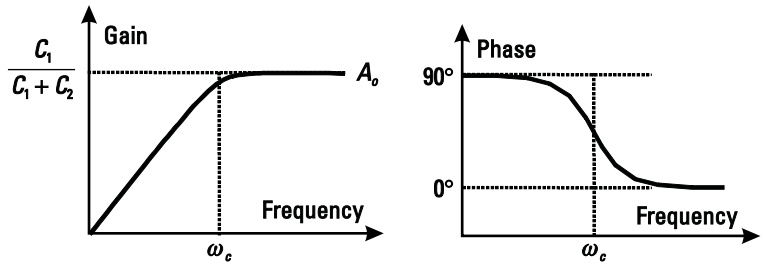

Figure 9.12 Plot of capacitive feedback frequency response.

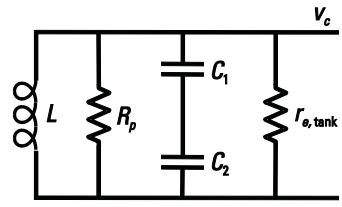

Figure 9.13 Z_{tank} using transformation of capacitive feedback divider.

for the Colpitts common base oscillator, and

$$r_{e,\text{tank}} = \left(1+\frac{C_1}{C_2}\right)^2 r_e \tag{9.24}$$

for the common collector oscillator. The resulting transformed circuit as seen by the resonator is shown in Figure 9.13.

Therefore, in order to get the maximum effect of the impedance transformation, it is necessary to make C_2 large and C_1 small in the case of the common-base circuit and vice versa for the common-collector circuit. However, one must keep in mind that the equivalent series capacitor nominally sets the resonance frequency according to

$$\omega = \frac{1}{\sqrt{LC_T}} = \sqrt{\frac{C_1+C_2}{LC_1C_2}} \tag{9.25}$$

Example 9.1: Capacitor Ratio

A common-base Colpitts oscillator with a resonance at 1.125 GHz using an on-chip inductor is required. Explore the role of the capacitor ratio on the emitter resistance transformation assuming that the largest available capacitor is 10 pF and the largest available inductor is 10 nH. Assume that the current through the transistor is set at 1 mA.

Solution:
Resonance frequency is given by

$$\omega = \frac{1}{\sqrt{LC_T}} = \sqrt{\frac{C_1+C_2}{LC_1C_2}}$$

and $r_{e,\text{tank}}$ is given by

$$r_{e,\text{tank}} = \left(1+\frac{C_2}{C_1}\right)^2 r_e$$

Large $r_{e,\text{tank}}$ is desired to reduce loss and minimize noise. To achieve this, it is advantageous for C_2 to be bigger than C_1. However, there will be a practical limit to the component values realizable on an integrated circuit. With 10 pF and 10 nH as the upper limits for capacitors and inductors on chip, Table 9.1 shows some of the possible combinations of L, C_1, and C_2 to achieve a frequency of 1.125 GHz.

Table 9.1 Inductor and Capacitor Values to Realize Oscillator at 1.125 GHz

L (nH)	C_T *(pF)*	C_1 *(pF)*	C_2 *(pF)*	*Resonance Frequency (GHz)*	*Impedance Transformation*	$r_{e,tank}$
10	2	2.5	10	1.125	25	625
8	2.5	3.333	10	1.125	16	400
8	2.5	3.75	7.5	1.125	9	225
6.667	3	4.5	9	1.125	9	225

Here, it can be seen that a transformation of 25 is about as high as is possible at this frequency and with the specified component limits as shown in the first row. Note that L and C_2 are still on the high side, indicating that designing an oscillator at this frequency with an impedance transformation of 25 is quite challenging. If the transformation can be reduced to 16 or 9, then a number of other choices are possible, as shown in the table. Note that at 1 mA, the emitter resistance is about 25 . Multiplying by 9 or 25 results in 225 or 625 . For a typical 10-nH inductor at 1.125 GHz, the equivalent parallel resistance might be 300 for a Q of 4.243. Even with this low inductor Q, $r_{e,\text{tank}}$ degrades the Q significantly. In the best case with a transformation of 25, the Q is reduced to less than 3.

Example 9.2: Frequency Shift

If an oscillator is designed as in Example 9.1 with a 10-nH inductor with Q of 4.243, and it is operated at 2.21 times the corner frequency of the highpass feedback network, then what is the expected frequency shift?

Solution:
By (9.17):

$$\omega_{\text{osc}} = \omega_o\sqrt{1+\frac{1}{4.243\times 2.21}} = 1.05\omega_o$$

so the frequency will be high by about 5%. This effect, due to the phase shift in the feedback path, is quite small and in practice can usually be neglected compared to the downward frequency shift due to parasitics and nonlinearities.

9.6.3 Oscillator Open-Loop Analysis

For this analysis, we redraw the Colpitts common-base oscillator with the loop broken at the emitter, as shown in Figure 9.14. Conceptually, one can imagine applying a small-signal voltage at v_e and measuring the loop gain at v_e . Since v_c is the output of the oscillator, we can define forward gain as v_c/v_e and feedback gain as v_e/v_c.

Forward Gain

The forward gain is:

$$H_1(s) = \frac{v_c}{v_e} = g_m Z_{\text{tank}} \tag{9.26}$$

where Z_{tank} is defined in Figure 9.15 and has the following transfer function:

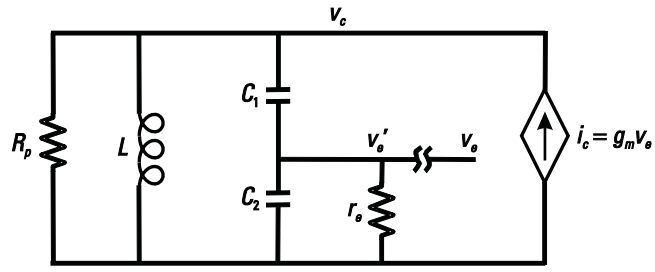

Figure 9.14 Feedback analysis of a Colpitts common-base oscillator.

$$Z_{\text{tank}} = Z_L \parallel R_p \parallel Z_{\text{FB}} = \frac{1}{\dfrac{1}{j\omega L} + \dfrac{1}{R_p} + \dfrac{j\omega C_1(1 + j\omega r_e C_2)}{1 + j\omega r_e(C_1 + C_2)}} \tag{9.27}$$

$$Z_{\text{tank}} = \frac{j\omega L R_p(1 + j\omega r_e(C_1 + C_2))}{(R_p + j\omega L)(1 + j\omega r_e(C_1 + C_2)) + j\omega L R_p \times j\omega C_1(1 + j\omega r_e C_2)} \tag{9.28}$$

Feedback Gain

The feedback circuit is just a highpass filter as described in the previous section and has the following transfer function:

$$H_2(j\omega) = \frac{v_e}{v_c} = \frac{j\omega r_e C_1}{1 + j\omega r_e(C_1 + C_2)} \tag{9.29}$$

Loop Gain Expression

This can be solved as follows:

$$A = H_1 H_2 = \frac{g_m \times j\omega r_e C_1 \times j\omega L R_p}{(R_p + j\omega L)(1 + j\omega r_e(C_1 + C_2)) + j\omega L R_p \times j\omega C_1(1 + j\omega r_e C_2)} \tag{9.30}$$

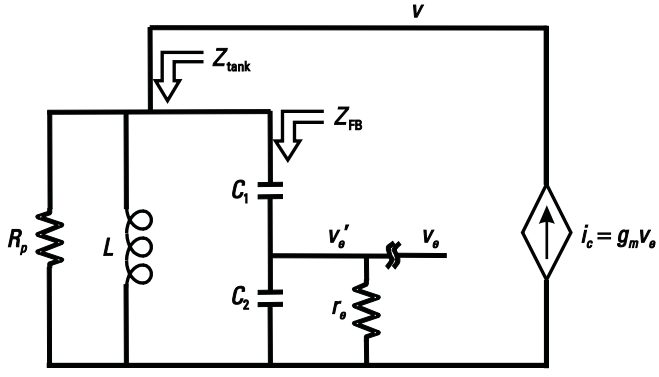

Figure 9.15 Definition of Z_{tank} and Z_{FB} in oscillator small-signal model.

Gathering terms:

$$A = H_1H_2 = \frac{\omega^2 \times g_m r_e C_1 L R_p}{B} \tag{9.31}$$

where

$$B = (\ j\omega^3) r_e C_1 C_2 L R_p + (-\omega^2)[r_e(C_1 + C_2)L + R_p C_1 L] + j\omega[r_e R_p (C_1 + C_2) + L] + R_p$$

To determine oscillating conditions, (9.31) can be set equal to 1, and then the real and imaginary terms can be solved independently. The real part, which includes the even-order terms, is set equal to 1, and this sets the condition for gain. The result can be shown to be the same as for the closed-loop analysis as done previously with final gain expression given by

$$g_m = \frac{\omega(C_1 + C_2)}{Q_L} \tag{9.32}$$

The imaginary part, which is defined by the odd-order terms, is set equal to zero. This is equivalent to setting the phase equal to zero. The result, again, is equal to the previous derivation with the result

$$\omega = \omega_o \sqrt{1 + \frac{1}{Q_L \omega / \omega_c}} \tag{9.33}$$

9.6.4 Simplified Loop Gain Estimates

To gain understanding, to explain this simple result, and to provide advice on how to do the design, in this section appropriate simplifications and approximations will be made by making use of the results shown in Section 9.6.2. As in the previous section, two expressions are written: one for the feedforward gain and one for the feedback gain.

If we assume we are operating above the capacitive feedback highpass corner frequency, then the feedback gain is given by

$$\frac{v_e}{v_c} = \frac{C_1}{C_1 + C_2} \tag{9.34}$$

Under these conditions, it can be seen that the capacitive voltage divider is a straight voltage divider with no phase shift involved. The loop gain can be seen to be

$$H_1H_2 = \frac{g_m}{Y_{\text{tank}}} \frac{C_1}{C_1 + C_2} = \frac{C_1}{C_1 + C_2} \div \times \frac{g_m}{\dfrac{1}{R_p} + \dfrac{1}{r_{e,\text{tank}}} + j\omega C_T \quad \dfrac{j}{\omega L}} \tag{9.35}$$

This can be set equal to 1 and solved for oscillating conditions. The imaginary terms cancel, resulting in the expected expression for resonant frequency:

$$\omega_o = \sqrt{\frac{1}{C_T L}} \tag{9.36}$$

The remaining real terms can be used to obtain an expression for the required g_m:

$$g_m = \frac{1}{R_p} + \frac{1}{r_{e,\text{tank}}} \div \times \frac{C_1 + C_2}{C_1} \div \tag{9.37}$$

where C_T, as before, is the series combination of C_1 and C_2. This final expression can be manipulated to show it is equal to (9.18) and (9.32) and is here repeated:

$$g_m = \frac{\omega(C_1 + C_2)}{Q_L} \tag{9.38}$$

Here we can see that the transistor transconductance makes up for losses in the resistors R_p and $r_{e,\text{tank}}$. Since they are in parallel with the resonator, we would like to make them as large as possible to minimize the loss (and the noise). We get large R_p by having large inductor Q, and we get large $r_{e,\text{tank}}$ by using a large value of capacitive transformer (by making C_2 bigger than C_1). Note that, as before, the value of g_m as specified in (9.37) or (9.38) is the value that makes loop gain equal to 1, which is the condition for marginal oscillation. To guarantee startup, loop gain is set greater than 1 or g_m is set greater than the value specified in the above equations.

Note in (9.38) r_e seems to have disappeared; however, it was absorbed by assuming that $g_m = 1/r_e$.

9.7 Negative Resistance Generated by the Amplifier

We have already talked about oscillators as a feedback system, but it is also possible to analyze them in terms of negative resistance. Negative resistance is generated by any circuit where, through feedback, an increase in voltage causes a decrease in current flow. If the negative resistance generated by the circuit is larger than the series losses of the circuit, then oscillations will persist. In the next few sections, we will explicitly derive formulas for how much negative resistance is generated by each type of oscillator.

9.7.1 Negative Resistance of the Colpitts Oscillator

In this section, an expression for the negative resistance of the oscillators will be derived. Consider first the common-base Colpitts configuration with the negative resistance portion of the circuit replaced by its small-signal model shown in Figure 9.16. Note that v_π and the current source have both had their polarity reversed for convenience.

An equation can be written for v_π in terms of the current flowing through this branch of the circuit.

$$i_i + g_m v_\pi = j\omega C_2 v_\pi + \frac{v_\pi}{r_e} \tag{9.39}$$

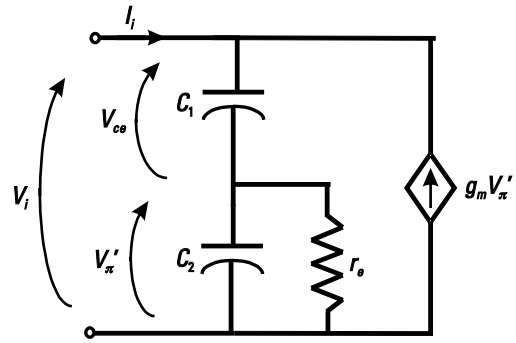

Figure 9.16 Small-signal model for the Colpitts common-base negative resistance cell.

This can be solved for v_π noting that $g_m \quad 1/r_e$.

$$v_\pi = \frac{i_i}{j\omega C_2} \tag{9.40}$$

Another equation can be written for v_{ce}.

$$v_{\text{ce}} = \frac{i_i + g_m v_\pi}{j\omega C_1} \tag{9.41}$$

Substituting (9.40) into (9.41) gives:

$$v_{ce} = \frac{1}{j\omega C_1} \div \; i_i + \frac{g_m i_i}{j\omega C_2} \div \tag{9.42}$$

Now using (9.40) and (9.42) and solving for $Z_i = v_i/i_i$ with some manipulation,

$$Z_i = \frac{v_i}{i_i} = \frac{v_\pi + v_{\text{ce}}}{i_i} = \frac{1}{j\omega C_1} + \frac{1}{j\omega C_2} \quad \frac{g_m}{\omega^2 C_1 C_2} \tag{9.43}$$

this is just a negative resistor in series with the two capacitors. Thus, a necessary condition for oscillation in this oscillator is

$$r_s < \frac{g_m}{\omega^2 C_1 C_2} \tag{9.44}$$

where r_s is the equivalent series resistance of the resonator. It will be shown in Example 9.4 that the series negative resistance is maximized for a given fixed total series capacitance when $C_1 = C_2$. An identical expression to (9.44) can be derived for the Colpitts common-collector circuit.

9.7.2 Negative Resistance for Series and Parallel Circuits

Equation (9.43) shows the analysis results for the oscillator circuit shown in Figure 9.17 when analyzed as an equivalent series circuit of C_1, C_2, and R_{neg}. Since the

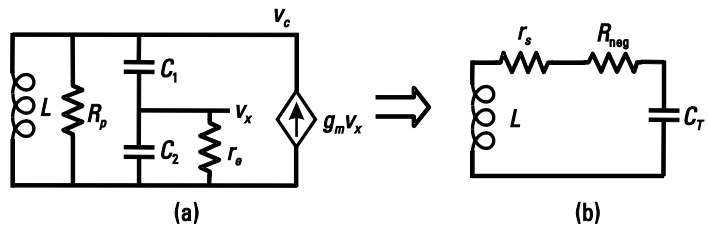

Figure 9.17 (a) Colpitts oscillator circuit and (b) equivalent series model.

resonance is actually a parallel one, the series components need to be converted back to parallel ones. However, if the equivalent Q of the RC circuit is high, the parallel capacitor, C_p will be approximately equal to the series capacitor C_s, and the above analysis is valid. Even for low Q, these simple equations are useful for quick calculations.

Example 9.3: Negative Resistance for Series and Parallel Circuits

Assume that, as before, L=10 nH, r_p=300 , C_1=2.5 pF, C_2=10 pF, and the transistor is operating at 1 mA, or $r_e = 25$ and g_m=0.04 A/V. Using negative resistance, determine the oscillator resonant frequency and apparent frequency shift.

Solution:
As before, C_T is 2 pF, and

$$\omega = \frac{1}{\sqrt{LC_T}} = \frac{1}{\sqrt{10\text{ nH}\quad 2\text{ pF}}} = 7.07\text{Grad/s}$$

or frequency $f_o = 1.1254$ GHz. The series negative resistance is equal to

$$r_{neg} = \frac{g_m}{\omega^2 C_1 C_2} = \frac{0.04}{(7.07107\text{ GHz})^2 \times 2.5\text{ pF} \times 10\text{ pF}} = 32.0$$

Then Q_{RC} can be calculated from

$$Q_{\text{RC}} = \frac{1}{\omega r_{neg} C_T} = \frac{1}{7.07107\text{ GHz} \times 32 \times 2\text{ pF}} = 2.2097$$

$$r_{\text{par}} = r_{neg}(1 + Q_{\text{RC}}^2) = 32(1 + 2.2097^2) = 188$$

We note that the parallel negative resistance is smaller in magnitude than the original parallel resistance, indicating that the oscillator should start up successfully.

The above is sufficient for a hand calculation; however, to complete the example, the equivalent parallel capacitance can be determined to be

$$C_{\text{par}} = \frac{C_T}{1 + 1/Q_{\text{RC}}^2} = \frac{2\text{ pF}}{1 + 1/2.2097^2} = 1.66\text{ pF}$$

This results in a new resonator resonant frequency of

$$\omega = \frac{1}{\sqrt{LC_{\text{par}}}} = \frac{1}{\sqrt{10\text{ nH}\quad 1.66\text{ pF}}} = 7.76150\text{ Grad/s}$$

This is a frequency of 1.2353 GHz, which is close to a 10% change in frequency. The oscillating frequency is determined by resonance of the loop, which in this case results in a 5% change in frequency as seen in Example 9.2. This discrepancy, which can be verified by a simulation of the original circuit of Figure 9.17(a), is due to the phase shift in the nonideal capacitive feedback path. Why is there a 10% frequency shift when a previous calculation showed a 5% frequency shift? The short answer is that the equation for frequency offset was derived for the marginal condition for oscillation. This would have a smaller value for g_m hence a smaller value for series negative resistance (in the marginal case it should exactly match the positive resistance). With a smaller negative resistance, the parallel capacitance will be larger; hence, the frequency will be lower. In fact the original frequency offset calculation should also be adjusted with the new larger value for r_e (assuming r_e is equal to the $1/g_m$). As a result, the actual frequency is 7.27607 Grad/sec for both approaches, which is about 3%.

While calculating frequency shifts and explaining them is of interest to academics, it is suggested that for practical designs, the simple calculations be used since parasitics and nonlinear effects will cause a downward shift of frequency. Further refinement should come from a simulator.

9.7.3 Negative Resistance Analysis of $-G_m$ Oscillator

The analysis of the negative resistance amplifier, shown in Figure 9.8, and the more common differential form in Figure 9.21(c) is somewhat different. The small-signal equivalent model for this circuit for the bipolar transistors is shown in Figure 9.18. The CMOS circuit would be identical except that v_π would be replaced with v_{gs}. Note that one transistor has had the normal convention reversed for v_π.

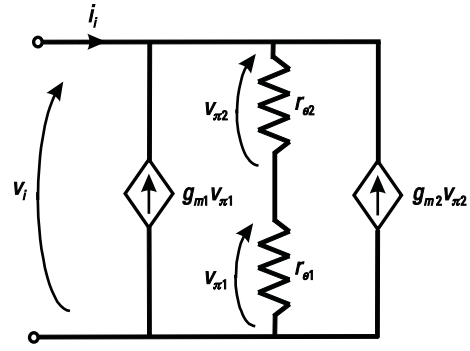

Figure 9.18 Small-signal equivalent model for the negative resistance cell in the negative resistance oscillator.

An expression for the current that flows into the circuit can be written as follows:

$$i_i = \frac{v_i}{r_{e1} + r_{e2}} \quad g_{m1}v_{\pi 1} \quad g_{m2}v_{\pi 2} \tag{9.45}$$

Now, if it is assumed that both transistors are biased identically, then $g_{m1} = g_{m2}$, $r_{e1} = r_{e2}$, $v_{\pi 1} = v_{\pi 2}$, and the equation can be solved for $Z_i = v_i/i_i$.

$$Z_i = \frac{2}{g_m} \tag{9.46}$$

Thus, in this circuit, a necessary condition for oscillation is that

$$g_m > \frac{2}{R_p} \tag{9.47}$$

where R_p is the equivalent parallel resistance of the resonator.

Example 9.4: Minimum Current for Oscillation

An oscillator is to oscillate at 3 GHz. Using a 5-nH inductor with $Q = 5$ and assuming no other loading on the resonator, determine the minimum current required to start the oscillations if a Colpitts oscillator is used or if a $-G_m$ oscillator is used.

Solution:
Ignoring the effect of the losses on the frequency of oscillation, we can determine what total resonator capacitance is required.

$$C_{\text{total}} = \frac{1}{\omega_{\text{osc}}{}^2 L} = \frac{1}{(2\pi \times 3\ \text{GHz})^2\, 5\ \text{nH}} = 562.9\ \text{fF}$$

The total capacitance is also given by:

$$C_{\text{total}} = \frac{C_1 C_2}{C_1 + C_2}$$

Since C_{total} is fixed because we have chosen a frequency of oscillation, we can solve for C_2:

$$C_2 = \frac{C_1 C_{\text{total}}}{C_1 \quad C_{\text{total}}}$$

Now we can put this back into the negative resistance formula in (9.44):

$$r_{\text{neg}} = \frac{g_m}{\omega^2 C_1 C_2} = \frac{g_m}{\omega^2 C_1 C_{\text{total}}} \quad \frac{g_m}{\omega^2 C_1{}^2}$$

To find the minimum current, we find the maximum r_{neg} by taking the derivative with respect to C_1.

$$\frac{r_{\text{neg}}}{C_1} = \frac{g_m}{\omega^2 C_1{}^2 C_{\text{total}}} + \frac{2g_m}{\omega^2 C_1{}^3} = 0$$

This leads to

$$C_1 = 2C_{\text{total}}$$

which means that the maximum obtainable negative resistance is achieved when the two capacitors are equal in value and twice the total capacitance. In this case $C_1 = C_2 = 1.1258$ pF.

Now the loss in the resonator at 3 GHz is due to the finite Q of the inductor.

The series resistance of the inductor is

$$r_s = \frac{\omega L}{Q} = \frac{(2\pi 3\text{ GHz})5\text{ nH}}{5} = 18.85$$

Therefore, $r_{\text{neg}} = r_s = 18.85$. Noting that $g_m = I_c/v_T$, and once more making use of (9.44)

$$I_C = \omega^2 C_1 C_2 v_T r_{\text{neg}} = (2\pi \times 3\text{ GHz})^2 (1.1258\text{ pF})^2 (25\text{ mV})(18.85\ ) = 212.2\ \mu\text{A}$$

In the case of the $-G_m$ oscillator there is no capacitor ratio to consider.

The parallel resistance of the inductor is:

$$R_p = \omega LQ = (2\pi \times 3\text{ GHz})5\text{ nH}(5) = 471.2$$

Therefore $r_{\text{neg}} = r_p = 471.2$. Noting again that $g_m = I_c/v_T$

$$I_C = \frac{2v_T}{R_p} = \frac{2(25\text{ mV})}{471.2} = 106.1\ \mu\text{A}$$

Thus, we can see from this example that a $-G_m$ oscillator can start with half as much collector current in each transistor as a Colpitts oscillator under the same loading conditions.

9.8 Comments on Oscillator Analysis

It has been shown that closed-loop analysis agrees exactly with the open-loop analysis. It can also be shown that analysis by negative resistance produces identical results. This analysis can be extended. For example, in a negative resistance oscillator, it is possible to determine if oscillations will be stable as shown by Kurokawa [1], with detailed analysis shown by [2]. However, what does it mean to have an exact analysis? Does this allow one to predict the frequency exactly? The answer is no. Even if one could take into account RF model complexities including parasitics, temperature, process, and voltage variations, the nonlinearities of an oscillator would still change the frequency. These nonlinearities are required to limit the amplitude of oscillation, so they are a built-in part of an oscillator. Fortunately, for a

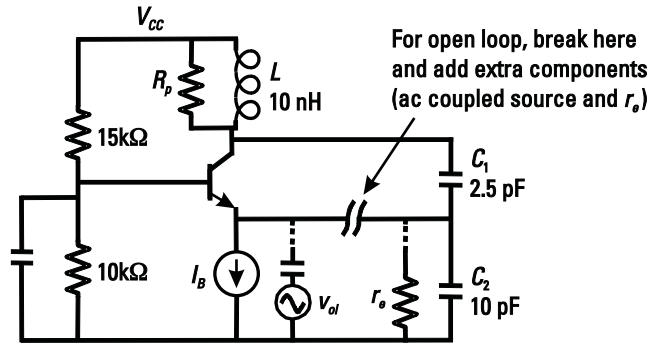

Figure 9.19 Circuit for oscillator simulations.

well-designed oscillator, the predicted results will give a reasonable estimate of the performance. Then to refine the design, it is necessary to simulate the circuit.

Example 9.5: Oscillator Frequency Shifts and Open-Loop Gain

Explore the predicted frequency with the actual frequency of oscillation by doing open-loop and closed-loop simulation of an oscillator. Compare the results to the simple formula. This example can also be used to explore the amplitude of oscillation and its relationship to the open-loop gain.

Solution:
For this example, the previously found capacitor and inductor values are used in the circuit shown in Figure 9.19.

Loop gain can be changed by adjusting g_m, or the tank resistance R_p. Both will also affect frequency somewhat. R_p will affect ω_c through Q_L and g_m will affect ω_c indirectly, since $r_e = 1/g_m$. In this case, we varied both R_p and g_m. Results are plotted in Figure 9.20.

It can be seen from Figure 9.20(a) that the open-loop simulations consistently predict higher oscillating frequencies than the closed-loop simulations. Thus, nonlinear behavior results in the frequency being decreased. We note that the initial frequency estimate using the inductor and capacitor values and adding an estimate for the parasitic capacitance results in a good estimate of final closed-loop oscillating frequency. In fact, this estimate of frequency is better than the open-loop small-signal prediction of frequency. It can also be seen from Figure 9.20(b) that output signal amplitude is related to the open-loop gain, and as expected, as gain drops to 1, or less, the oscillations stop.

So how does one decide on the oscillator small-signal loop gain? In a typical RF integrated oscillator, a typical starting point is to choose a small-signal loop (voltage) gain of about 1.4 to 2 (or 3 to 6 dB), then the current is swept to determine the minimum phase noise. Alternatively, one might design for optimal output power; however, typically output buffers are used to obtain the desired output power and most oscillators are designed for best phase noise performance. In traditional negative resistance oscillators, analysis has shown that small-signal open-loop voltage

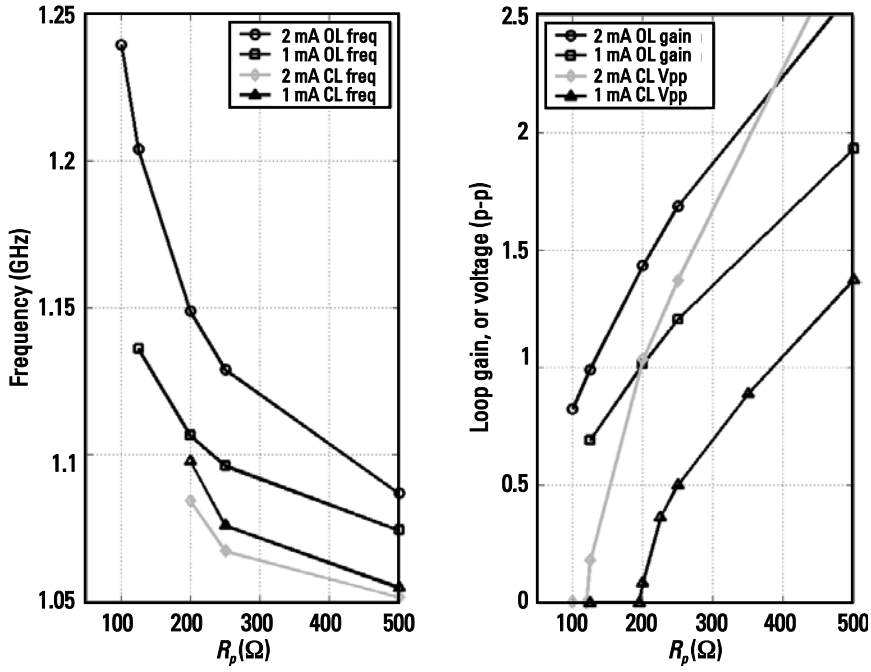

Figure 9.20 Plot of oscillator performance versus tank resistor: (a) open-loop and closed-loop frequency, and (b) loop gain and oscillation amplitude.

gain of 3 is optimal for output power [2]. Fortunately, this is close to the optimum value for phase noise performance.

9.9 Basic Differential Oscillator Topologies

The three main oscillators discussed so far can be made into differential circuits. The basic idea is to take two single-ended oscillators and place them back to back. The nodes in the single-ended circuits, which were previously connected to ground, in the differential circuit are tied together forming an axis of symmetry down the center of the circuit. The basic circuits with biasing are shown in Figure 9.21. Note that in the conversion from single-ended to differential, the frequency is changed unless total inductance and capacitance L and C are either kept as L and C or converted to $2L$ and $C/2$.

9.10 A Modified Common-Collector Colpitts Oscillator with Buffering

One problem with oscillators is that they must be buffered in order to drive a low impedance. Any load that is a significant fraction of the R_p of the oscillator would lower the output swing and increase the phase noise of the oscillator. It is common to buffer oscillators with a stage such as an emitter follower or emitter-coupled pair.

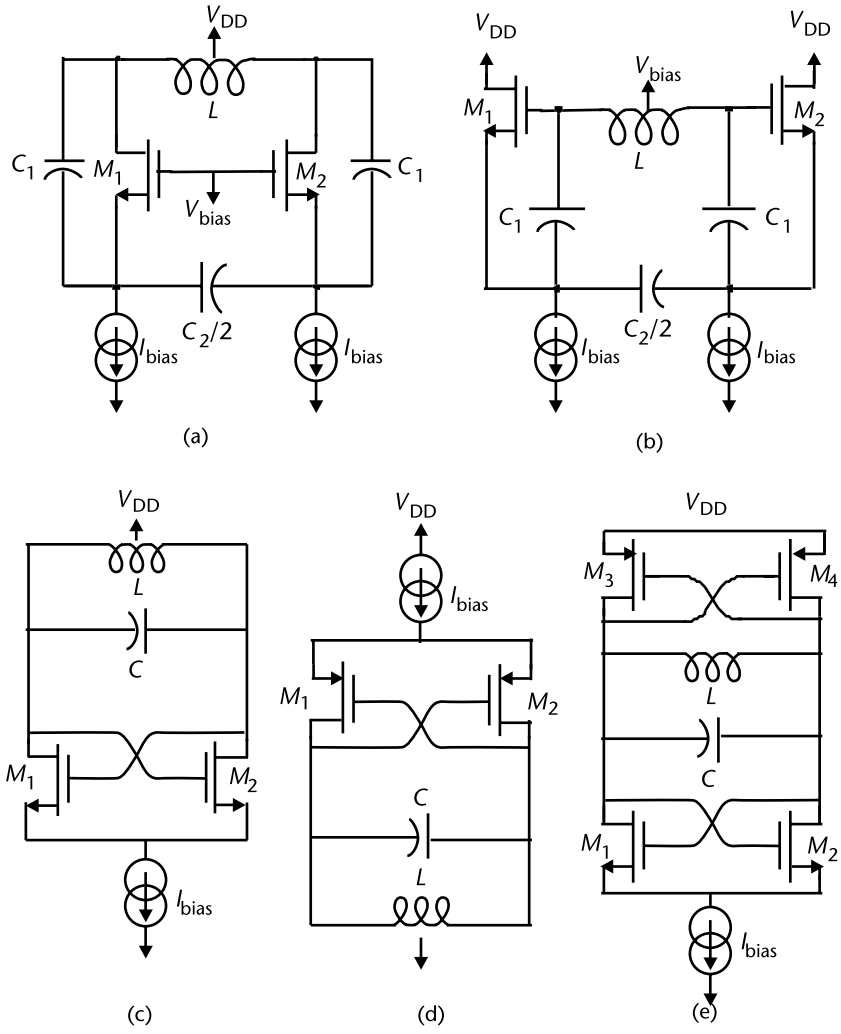

Figure 9.21 Basic differential oscillator: (a) Colpitts common-gate, (b) Colpitts common-drain, (c) G_m oscillator, NMOS only, (d) G_m oscillator, PMOS only, and (e) G_m oscillator, complementary.

These stages add complexity and require current. One design that gets around this problem is shown in Figure 9.22. Here, the common-collector oscillator is modified slightly by the addition of resistors placed in the collector [3, 4]. The output is then taken from the collector. Since this is a high impedance node, the oscillator's resonator is isolated from the load without using any additional transistors or current. However, the addition of these resistors will also reduce the headroom available to the oscillator.

9.11 Several Refinements to the G_m Topology Using Bipolar Transistors

Several refinements can be made to the $-G_m$ oscillator to improve its performance. In the version already presented in Figure 9.21(c), the transistors' gates and drains

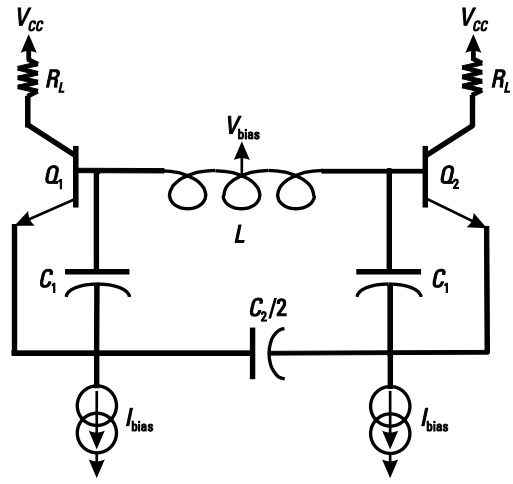

Figure 9.22 Modified Colpitts common-collector oscillator with self-buffering.

are at the same dc voltage. If this circuit is made with bipolar transistors, the maximum voltage swing that can be obtained is about 0.8V. That is to say, the voltage on one side of the resonator drops about 0.4V, while the voltage on the other side of the oscillator rises by about 0.4V. This means that the collector would be about 0.8V below the base and the transistor goes into saturation. In order to get larger swings out of this topology, we must decouple the base from the collector. One common way to do this is with capacitors. This improved oscillator is shown in Figure 9.23. The bases have to be biased separately now, of course. Typically, this is done by placing resistors in the bias line. These resistors have to be made large to prevent loss of signal at the base. However, these resistors can be a substantial source of noise.

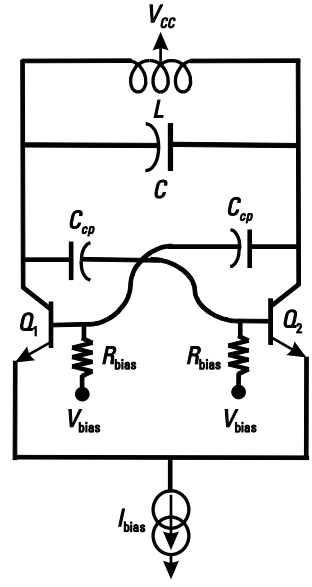

Figure 9.23 G_m oscillator with capacitive decoupling of the bases.

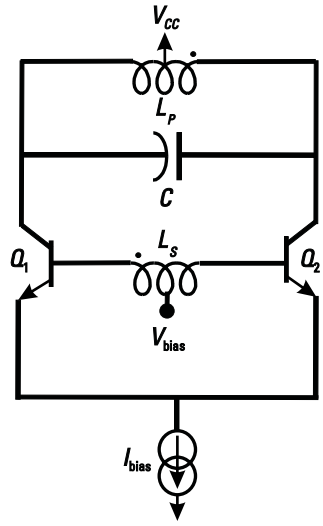

Figure 9.24 G_m oscillator with inductive decoupling of the bases.

Another variation on this topology is to use a transformer instead of capacitors to decouple the collectors from the bases, as shown in Figure 9.24 [5]. Since the bias can be applied through the center tap of the transformer, there is no longer a need for the RF blocking resistors in the bias line. Also, if a turns ratio of greater than unity is chosen, there is the added advantage that the swing on the base can be much smaller than the swing on the collector, helping to prevent transistor saturation.

Another modification that can be made to the $-G_m$ oscillator is to replace the high-impedance current source connected to the emitters of Q_1 and Q_2 in Figure 9.24 with a resistor. Since the resistor is not a high-impedance source, the bias current will vary dynamically over the cycle of the oscillation. In fact, the current will be highest when the oscillator voltage is at its peaks and lowest during the zero crossings of the waveform. Since the oscillator is most sensitive to phase noise during the zero crossings, this version of the oscillator can often give very good phase noise performance. This oscillator is shown in Figure 9.25(a).

A brief circuit description will now be provided. The circuit must be dc-biased at some low current. As the oscillation begins, the voltage rises on one side of the resonator and one transistor starts to turn off while the other starts to turn on harder and draw more current. As the transistor draws more current, more current flows through R_{tail}, and thus the voltage across this resistor starts to rise. This acts to reduce the v_{BE} of the transistor which acts as feedback to limit the current at the top and bottom of the swing. The collector waveforms are shown conceptually in Figure 9.26. Since the current is varying dynamically over a cycle, and since the resistor R_{tail} does not require as much headroom as a current source, this allows a larger oscillation amplitude for a given power supply.

An alternative to the resistor R_{tail} is to use a noise filter in the tail as shown in Figure 9.25(b) [6]. Note this filter may also benefit CMOS implementations. While the use of the inductor does require more chip area, its use can lead to a very low-noise bias, leading to low-phase-noise designs. Another advantage to using this

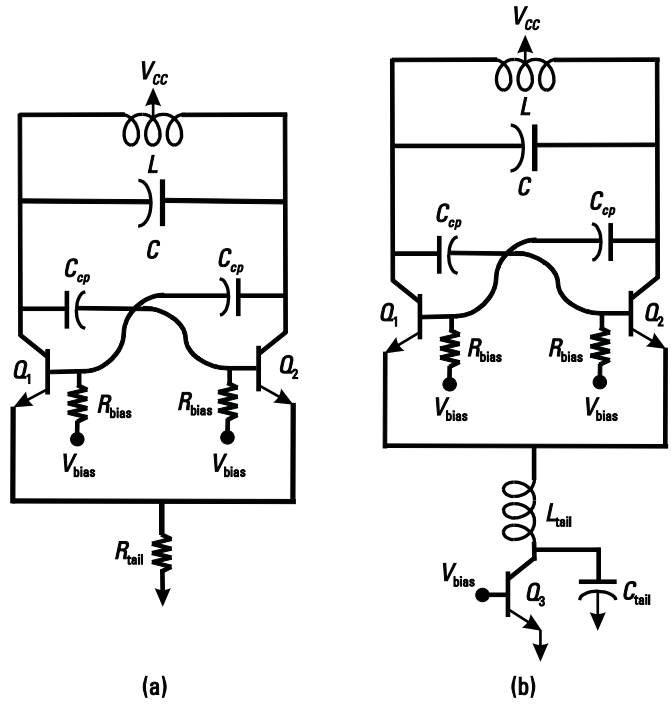

Figure 9.25 G_m oscillator with (a) resistive tail current source or (b) current source noise filter.

noise filter is that before startup, the transistor Q_3 can be biased in saturation, because during startup the second harmonic will cause a dc bias shift at the collector of Q_3, pulling it out of saturation and into the active region. Also, since the second harmonic cannot pass through the inductor L_{tail}, there is no "ringing" at the collector of Q_3, further reducing its headroom requirement.

9.12 The Effect of Parasitics on the Frequency of Oscillation

The first task in designing an oscillator is to set the frequency of oscillation and hence set the value of the total inductance and capacitance in the circuit. To increase

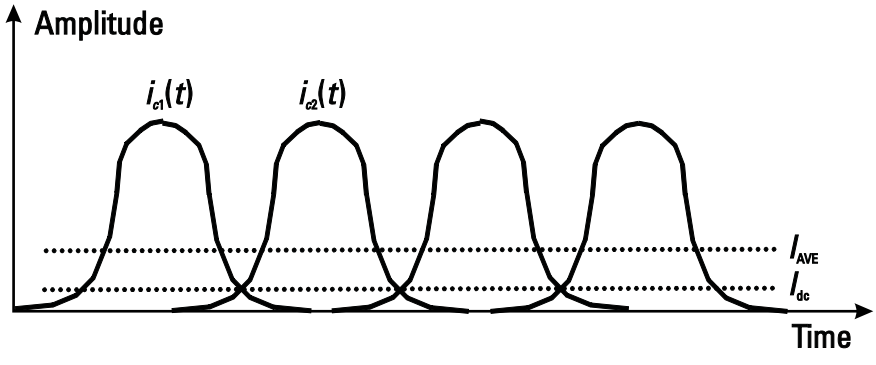

Figure 9.26 G_m oscillator with resistive tail collector currents.

output swing, it is usually desirable to make the inductance as large as possible (this will also make the oscillator less sensitive to parasitic resistance). However, it should be noted that large monolithic inductors suffer from limited Q. In addition, as the capacitors become smaller, their value will be more sensitive to parasitics. The frequency of oscillation, for the Colpitts common base oscillator as shown in Figure 9.9(a), taking into account transistor parasitics, is given by

$$\omega_{\text{osc}} \quad \frac{1}{\sqrt{L \ \frac{C_1C_2 + C_1C_\pi}{C_1 + C_2 + C_\pi} + C_{\mu\div}}} \tag{9.48}$$

For the Colpitts common-collector oscillator, as shown in Figure 9.9(b), the frequency is given by

$$\omega_{\text{osc}} \quad \frac{1}{\sqrt{L \ \frac{C_1C_2 + C_2C_\pi}{C_1 + C_2 + C_\pi} + C_{\mu\div}}} \tag{9.49}$$

For the $-G_m$ oscillator, the frequency is given by

$$\omega_{\text{osc}} \quad \frac{1}{\sqrt{L \ \ 2C_\mu + \frac{C_\pi}{2} + C_\div}} \tag{9.50}$$

Note that in the case of the $-G_m$ oscillator, the parasitics tend to reduce the frequency of oscillation a bit more than with the Colpitts oscillator. For CMOS the results would be identical except that C_π would be replaced with C_{gs} and C_μ would be replaced with C_{gd}.

9.13 Large-Signal Nonlinearity in the Transistor

So far, the discussion of oscillators has assumed that the small-signal equivalent model for the transistor is valid. If this were true, then the oscillation amplitude would grow indefinitely, which is not the case. As the signal grows, nonlinearity will serve to reduce the negative resistance of the oscillator until it just cancels out the losses and the oscillation reaches some steady-state amplitude. The source of the nonlinearity is typically the transistor itself.

Usually the transistor is biased somewhere in the active region for bipolar or the saturation region for CMOS. At this operating point, the transistor will have a particular g_m. However, as the voltage swing starts to increase during startup, the instantaneous g_m will start to change over a complete cycle. The transistor may even start to enter the saturation region for bipolar or the triode region for CMOS at one end of the swing and the cutoff region at the other end of the voltage swing. Which of these effects starts to happen first depends on the biasing of the transistor. Ultimately, a combination of all effects may be present. Eventually, with increasing signal amplitude, the effective g_m will decrease to the point where it just compensates for the losses in the circuit and the amplitude of the oscillator will stabilize.

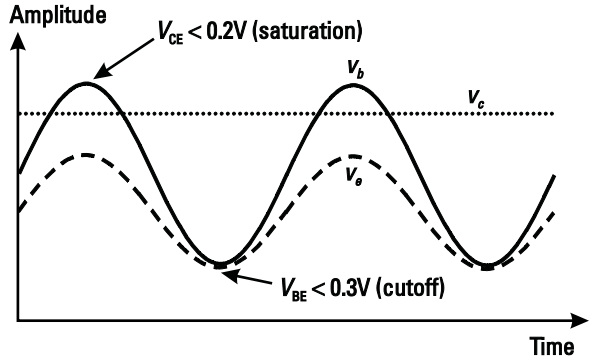

Figure 9.27 Waveforms for a common-collector oscillator that is heavily voltage-limited.

The saturation/triode and cutoff linearity constraint will also put a practical limit on the maximum power that can be obtained from an oscillator. After reaching this limit, increasing the bias current will have very little effect on the output swing. Although increasing the current causes the small-signal g_m to rise, this just tends to "square up" the signal rather than to increase its amplitude.

Looking at the common-base or common-collector Colpitts oscillators as shown in Figure 9.9, it can be seen how this effect works on the circuit waveforms in Figure 9.27 and Figure 9.28. In the case of the common-base circuit, when v_c is at the bottom of its swing, v_{ce} tends to be very small, causing the base collector junction to be forward-biased. This also tends to make v_{be} quite large. These two conditions together cause the transistor to go into saturation. When v_c reaches the top of its swing, v_{be} gets very small and this drives the oscillator into cutoff. A similar argument can be made for the common-collector circuit, except that it enters cutoff at the bottom of its swing and saturation at the top of its swing.

9.14 Bias Shifting During Startup

Once the oscillator starts to experience nonlinearity, harmonics start to appear. The even-ordered harmonics, if present, can cause shifts in bias conditions since they are

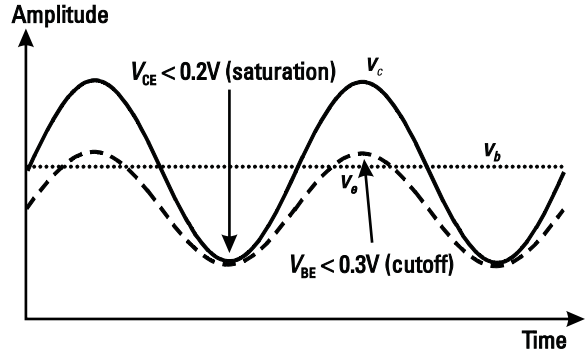

Figure 9.28 Waveforms for a common-base oscillator that is heavily voltage-limited.

not symmetric. They have no negative-going swing so they can change the average voltage or current at a node. Thus, they tend to raise the voltage at any node with signal swing on it, and after startup, bias conditions may shift significantly from what would be predicted by a purely dc analysis. For instance, the voltage at the emitter of the common-base or common-collector Colpitts oscillators will tend to rise. Another very good example of this is the $-G_m$ oscillator with resistive tail as shown in Figure 9.25. The node connected to R_{tail} is a virtual ground at the fundamental frequency; however, there is strong second-harmonic content on this node that tends to raise the average voltage level and the current through the oscillator after startup.

9.15 Colpitts Oscillator Amplitude

If the oscillator satisfies the conditions for oscillation, then oscillations will continue to grow until the transistor nonlinearities reduce the gain making the losses and the negative resistance of equal value.

For a quantitative analysis of oscillation amplitude, we first start with a transistor being driven by a large sinusoidal voltage, as shown in Figure 9.29. Note that in a real oscillator, a sinusoid driving the base is a good approximation, provided the resonator has a reasonable Q. This results in all other frequency components being filtered out and the voltage (although not the current) is sinusoidal even in the presence of strong nonlinearity.

It is assumed that the transistor is being driven by a large voltage, so it will only be on for a very small part of the cycle, during which time it produces a large pulse of current. However, regardless of what the current waveform looks like, its average value over a cycle must still equal the bias current. Therefore,

$$\overline{i_c} = \frac{1}{T}\int_0^T i_c(t)dt = I_{\text{bias}} \tag{9.51}$$

The part of the current at the fundamental frequency of interest can be extracted by multiplying by a cosine at the fundamental and integrating.

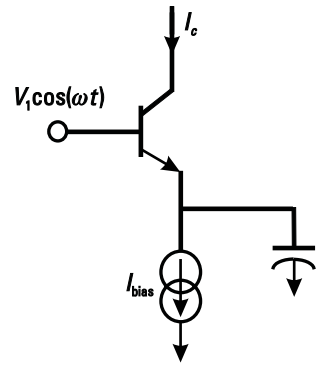

Figure 9.29 Transistor driven by a large sinusoidal voltage source.

$$i_{\text{fund}} = \frac{2}{T}\int_0^T i_c(t)\cos(\omega t)dt \tag{9.52}$$

This can be solved by assuming a waveform for $i_c(t)$. However, solving this equation can be avoided by noting that the current pulses are very narrow, and therefore, the cosine can be approximated as unity and integration simplifies:

$$i_{\text{fund}} \quad \frac{2}{T}\int_0^T i_c(t)dt = 2I_{\text{bias}} \tag{9.53}$$

With this information, it is possible to define a large signal transconductance for the transistor given by

$$G_m = \frac{i_{\text{fund}}}{V_1} = \frac{2I_{\text{bias}}}{V_1} \tag{9.54}$$

Since $g_m = I_C/v_T$ and since G_m can never be larger than g_m, it becomes clear that this approximation is not valid if V_1 is less than $2v_T$.

We can now apply this to the case of the Colpitts common-collector oscillator as shown in Figure 9.9. We draw the simplified schematic replacing the transistor with the large-signal transconductance as shown in Figure 9.30. Note that we are using the T-model here for the transistor, so the current source is between collector and base.

We first note that the resonator voltage will be the bias current at the fundamental times the equivalent resonator resistance.

$$V_{\text{tank}} = 2I_{\text{bias}}R_{\text{total}} \tag{9.55}$$

Since R_{total} is the total differential resistance, this is peak differential voltage.

This resistance will be made up of the equivalent loading of all losses in the oscillator and the loading of the transconductor on the resonator.

The transconductor presents the impedance

$$\frac{1}{G_m}\left(\frac{C_1 + C_2}{C_2}\right)^2 = \frac{1}{G_m n^2} \tag{9.56}$$

where n is the equivalent impedance transformation ratio.

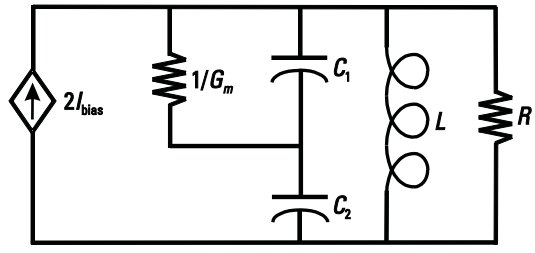

Figure 9.30 Colpitts common-collector oscillator with large-signal transconductance applied.

This is in parallel with all other losses in the resonator R_p:

$$R_{\text{total}} = R_p \,//\, \frac{1}{G_m n^2} = \frac{R_p}{1 + G_m n^2 R_p} \tag{9.57}$$

We can plug this back into the original expression:

$$V_{\text{tank}} = 2I_{\text{bias}} \frac{R_p}{1 + G_m n^2 R_p} \tag{9.58}$$

Now we also know that

$$G_m = \frac{2I_{\text{bias}}}{V_1} \tag{9.59}$$

and that

$$V_1 = \frac{C_2}{C_1 + C_2} V_{\text{tank}} = n V_{\text{tank}} \tag{9.60}$$

Therefore,

$$V_{\text{tank}} = 2I_{\text{bias}} \frac{R_p}{1 + \dfrac{2I_{\text{bias}}}{n V_{\text{tank}}} n^2 R_p} \tag{9.61}$$

$$V_{\text{tank}} = 2I_{\text{bias}} R_p \quad \frac{C_1}{C_1 + C_2} \div \tag{9.62}$$

A very similar analysis can be carried out for the common-base Colpitts oscillator shown in Figure 9.9 and it will yield the result

$$V_{\text{tank}} = 2I_{\text{bias}} R_p \quad \frac{C_2}{C_1 + C_2} \div \tag{9.63}$$

Note that it is often common practice to place some degeneration in the emitter of the transistor in a Colpitts design. This practice will tend to spread the pulses over a wider fraction of a cycle, reducing the accuracy of the above equation somewhat. However, it should still be a useful estimate of oscillation amplitude.

9.16 G_m Oscillator Amplitude

The determination of amplitude of the $-G_m$ oscillator will now be considered. The amplitude is dependent on whether the oscillator is voltage- or current-limited. If the oscillator is voltage-limited, typically, it is limited by the on-voltage of a switch, either to ground (for a bipolar or NMOS-only oscillator) or to V_{DD} (for PMOS-only oscillator). For an NMOS-only oscillator, the outputs are nominally at V_{DD}, with the maximum negative swing approaching ground and thus by symmetry, the

positive swing is expected to go to $2V_{DD}$ for a peak swing of V_{DD} per side, or peak swing of $2V_{DD}$ for a differential output. Similarly, for a PMOS-only oscillator, the outputs are nominally at ground potential, and pulled up by the PMOS switches towards V_{DD}. This peak positive swing of V_{DD} per side is matched by a peak negative swing approaching $-V_{DD}$, again for a peak differential swing of $2V_{DD}$. For a complementary circuit with both PMOS and NMOS switches, both sides of the inductor are connected either to ground or to V_{DD}, hence the maximum possible peak differential swing is V_{DD}. Note that bipolar transistors have a diode from base to collector, thus, bipolar negative G_m oscillators will voltage limit to $2V_{BE}$. For this reason, capacitors are often inserted between the collector and base of bipolar negative G_m oscillators to decouple the dc components. In such a case, biasing resistors need to be used to bias the gate. Since CMOS transistors do not have an equivalent diode (from gate to drain), such coupling capacitors are not needed in CMOS negative G_m oscillators.

For current-limited oscillators, as long as the voltage is high enough then the transistors can be treated as switches. Thus, each side will have current that switches between 0 and I_{bias}. A simple Fourier analysis can be used to show that the average value is $I_{\mathrm{bias}}/2$ and the peak fundamental value is $2I_{\mathrm{bias}}/\pi$. Thus, since the impedance per side is $R_p/2$ (current is flowing into the center tap of the inductor) the output voltage per side is:

$$v_{\mathrm{out}}\big|_{\mathrm{SE}} = \frac{R_p I_{\mathrm{bias}}}{\pi} \tag{9.64}$$

and the differential voltage is:

$$v_{\mathrm{out}}\big|_{\mathrm{DE}} = \frac{2R_p I_{\mathrm{bias}}}{\pi} \tag{9.65}$$

Note that this result can also be applied to a bipolar $-G_m$ oscillator.

With complementary transistors, the current is flowing through the full R_p in each direction; thus, the output voltage is twice as large:

$$v_{\mathrm{out}}\big|_{\mathrm{Comp\_DE}} = \frac{4R_p I_{\mathrm{bias}}}{\pi} \tag{9.66}$$

9.17 Phase Noise

A major challenge in most oscillator designs is to meet the phase noise requirements of the system. An ideal oscillator has a frequency response that is a simple impulse at the frequency of oscillation. However, real oscillators exhibit "skirts" caused by instantaneous jitter in the phase of the waveform. Noise that causes variations in the phase of the signal (distinct from noise that causes fluctuations in the amplitude of the signal) is referred to as phase noise. Because of amplitude limiting in integrated oscillators, typically AM noise is lower than FM noise. There are several major sources of phase noise in an oscillator, and they will be discussed next.

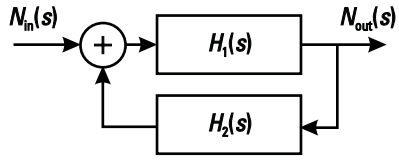

Figure 9.31 Feedback model of an oscillator used for phase-noise modeling.

9.17.1 Linear or Additive Phase Noise and Leeson's Formula

In order to derive a formula for phase noise in an oscillator, we will start with the feedback model of an oscillator as shown in Figure 9.31 [7].

From control theory, it is known that

$$\frac{N_{\text{out}}(s)}{N_{\text{in}}(s)} = \frac{H_1(s)}{1 \quad H(s)} \tag{9.67}$$

where $H(s) = H_1(s)\ H_2(\text{s})$. $H(s)$ can be written as a truncated Taylor series:

$$H(j\omega) \quad H(j\omega_o) + \quad \omega\frac{dH}{d\omega} \tag{9.68}$$

Since the conditions of stable oscillation must be satisfied, $H(j\omega_o) = 1$. We let $H_1(j\omega_o) = H_1$, where H_1 is a constant determined by circuit parameters.

Now (9.67) can be rewritten using (9.68) as

$$\frac{N_{\text{out}}(s)}{N_{\text{in}}(s)} = \frac{H_1}{\omega\dfrac{dH}{d\omega}} \tag{9.69}$$

Noise power is of interest here, so

$$\left|\frac{N_{\text{out}}(s)}{N_{\text{in}}(s)}\right|^2 = \frac{|H_1|^2}{(\ \omega)^2\left|\dfrac{dH}{d\omega}\right|^2} \tag{9.70}$$

This equation can now be rewritten using $H(\omega) = |H|e^{j\phi}$ and the product rule

$$\frac{dH}{d\omega} = \frac{d|H|}{d\omega}e^{j\phi} + |H|je^{j\ \phi}\frac{d\phi}{d\omega} \tag{9.71}$$

noting that the two terms on the right are orthogonal:

$$\left|\frac{dH}{d\omega}\right|^2 = \left|\frac{d|H|}{d\omega}\right|^2 + |H|^2\left|\frac{d\phi}{d\omega}\right|^2 \tag{9.72}$$

At resonance, the phase changes much faster than magnitude, and $|H| \quad 1$ near resonance. Thus, the second term on the right is dominant, and this equation reduces to

$$\left|\frac{dH}{d\omega}\right|^2 = \left|\frac{d\phi}{d\omega}\right|^2 \tag{9.73}$$

Now substituting (9.73) back into (9.70),

$$\left|\frac{N_{\text{out}}(s)}{N_{\text{in}}(s)}\right|^2 = \frac{|H_1|^2}{(\ \omega)^2\left|\frac{d\phi}{d\omega}\right|^2} \tag{9.74}$$

This can be rewritten again with the help of the definition of Q given in Chapter 5:

$$\left|\frac{N_{\text{out}}(s)}{N_{\text{in}}(s)}\right|^2 = \frac{|H_1|^2\,\omega_o{}^2}{4Q^2(\ \omega)^2} \tag{9.75}$$

In the special case for which the feedback path is unity, then $H_1 = H$, and since $|H| = 1$ near resonance it reduces to

$$\left|\frac{N_{\text{out}}(s)}{N_{\text{in}}(s)}\right|^2 = \frac{\omega_o{}^2}{4Q^2(\ \omega)^2} \tag{9.76}$$

Equation (9.76) forms the noise shaping function for the oscillator. In other words, for a given noise power generated by the transistor amplifier part of the oscillator, this equation describes the output noise around the tone.

Phase noise is usually quoted as an absolute noise referenced to the carrier power, so (9.76) should be rewritten to give phase noise as

$$\text{PN} = \frac{|N_{\text{out}}(s)|^2}{2P_S} = \frac{|H_1|\omega_o}{(2Q\ \omega)\div}^2 \frac{|N_{\text{in}}(s)|^2}{2P_S}\div \tag{9.77}$$

where P_S is the signal power of the carrier and noting that phase noise is only half the noise present. The other half is amplitude noise, which is of less interest. As well, in this approximation, conversion of amplitude noise to phase noise (also called AM to PM conversion) is ignored. This formula is known as Leeson's equation [8].

The one question that remains here is: What exactly is N_{in}? If the transistor and bias were assumed to be noiseless, then the only noise present would be due to the resonator losses. Since the total resonator losses are due to its finite resistance, which has an available noise power of kT, then

$$|N_{\text{in}}(\text{s})|^2 = kT \tag{9.78}$$

The transistors and the bias will add noise to this minimum. Note that since this is not a simple amplifier with a clearly defined input and output, it would not be appropriate to define the transistor in terms of a simple noise figure. Considering the bias noise in the case of the $-G_m$ oscillator, as shown in Figure 9.21(c), noise will come from the current source when the transistors are switched. If ρ is the fraction

of a cycle for which the transistors are completely switched, i_{nt} is the noise current injected into the oscillator from the biasing network during this time. During transitions, the transistors act like an amplifier, and thus collector shot noise i_{cn} from the resonator transistors usually dominates the noise during this time when the oscillator is built using bipolar transistors. The total input noise becomes

$$\left|N_{\text{in}}(s)\right|^2 \quad kT + \frac{i_{nt}^2 R_p}{2}\rho + i_{cn}{}^2 R_p(1 \quad \rho) \tag{9.79}$$

where R_p is the equivalent parallel resistance of the tank. Thus, we can define an excess noise factor for the oscillator as excess noise injected by noise sources other than the losses in the tank:

$$F = 1 + \frac{i_{nt}^2 R_p}{2kT}\rho + \frac{i_{cn}{}^2 R_p(1 \quad \rho)}{kT} \tag{9.80}$$

Note that this equation is too simplistic in that the oscillator is not equally sensitive to injected noise in all parts of the cycle. Because the oscillator is most sensitive to injected noise during the transitions, the term for noise during the fully switched times is much less important in the estimate of noise factor. Also note that as the Q of the tank increases, R_p increases and noise has more gain to the output; therefore, F is increased. Thus, while (9.76) shows a decrease in phase noise with an increase in Q, this is somewhat offset by the increase in F. If noise from the bias i_{nt} is filtered and if fast switching is employed, it is possible to achieve a noise factor close to unity in bipolar oscillators. For CMOS oscillators, ignoring the component during the fully switched times, the noise factor becomes

$$F = 1 + 4\gamma g_m R_p(1 \quad \rho) \tag{9.81}$$

Now (9.77) can be rewritten as

$$\text{PN} = \frac{\left|H_1\right|\omega_o}{(2Q \quad \omega)}\div^2 \quad \frac{FkT}{2P_S}\div \tag{9.82}$$

Note that in this derivation, it has been assumed that flicker noise is insignificant at the frequencies of interest. This may not always be the case, especially in CMOS designs. If ω_c represents the flicker noise corner where flicker noise and thermal noise are equal in importance, then (9.82) can be rewritten as

$$\text{PN} = \frac{\left|H_1\right|\omega_o}{(2Q \quad \omega)}\div^2 \quad \frac{FkT}{2P_S}\div \quad 1 + \frac{\omega_c}{\omega}\div \tag{9.83}$$

It can be noted that (9.83) predicts that noise will roll off at slopes of 30 or 20 dB/decade depending on whether flicker noise is important. However, in real life, at high-frequency offsets there will be a thermal noise floor. A typical plot of phase noise versus offset frequency is shown in Figure 9.32.

It is important to make a few notes here about the interpretation of this formula. Note that in the derivation of this formula, it has been assumed that the

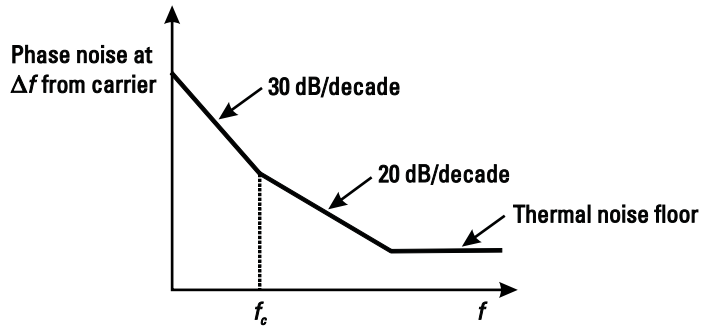

Figure 9.32 Phase noise versus frequency.

noise N_{IN} is injected into the resonator. Thus $|H_1|=|H|=1$ at the top of the resonator. However, at other points in the loop, the signal level is not the same as at the resonator. For instance, in the case of the common-base Colpitts oscillator, when looking at the midpoint between the two capacitors, then the signal is reduced by $C_2/(C_1 + C_2)$. A common mistake is to assume that, in this circuit, the phase noise at this point would be intrinsically worse than at the top of the resonator because the output power is lower by a factor $[C_2/(C_1 + C_2)]^2$. However, the noise is reduced by the same amount, leaving the phase noise at the same level at both points in the feedback loop. Note that this is only true for offset frequencies for which the noise is higher than the thermal noise floor.

Example 9.6: Phase Noise Limits

A sales representative for Simply Fabless Semiconductor Inc. has told a potential customer that Simply Fabless can deliver a 5-GHz receiver including an on-chip phase-locked loop (PLL). The VCO in the part is to run off a 1.8-V supply, consume no more than 1 mW of power, and deliver a phase noise performance of –105 dBc/Hz at 100-kHz offset. It has fallen on the shoulders of engineering to design this part. It is known that, in the technology to be used, the best inductor Q is 15 for a 3-nH device. Assume that capacitors or varactors will have a Q of 50. What is the likelihood that engineering will be able to deliver the part with the required performance to the customer?

Solution:
We will assume a $-G_m$ topology for this design and start with the assumption that the inductor and capacitive resistance are the only load on the device (we will ignore all other losses).

The $r_{p/L}$ of the inductor is

$$r_{p/L} = \omega LQ = 2\pi \times 5\ \text{GHz} \times 3\ \text{nH} \times 15 = 1,413.7$$

The capacitance in the design will be

$$C_{\text{total}} = \frac{1}{\omega_{\text{osc}}{}^2 L} = \frac{1}{(2\pi \times 5\ \text{GHz})^2\, 3\ \text{nH}} = 337.7\ \text{fF}$$

Thus, the parallel resistance due to the capacitor will be

$$r_{p/C} = \frac{Q}{\omega C_{\text{total}}} = \frac{50}{(2\pi \times 5\ \text{GHz}) \times 337.7\ \text{fF}} = 4{,}712.9$$

Thus, the equivalent parallel resistance of the resonator is 1,087.5 .

With a supply of 1.8V and a power consumption of 1 mW, the maximum current that the circuit can draw is 555.5 μA.

The peak voltage swing in the oscillator will be

$$V_{\text{tank}} = \frac{2}{\pi} I_{\text{bias}} R_p = \frac{2}{\pi}(555.5\ \mu\text{A})(1{,}087.5\ \ ) = 0.384\text{V}$$

This means that the oscillator will have an RF output power of

$$P = \frac{V_{\text{tank}}{}^2}{2R_p} = \frac{(0.384\text{V})^2}{2(1{,}087.5\ \ )} = 67.8\ \mu\text{W}$$

The Q of the oscillator will be

$$Q = R_P \sqrt{\frac{C_{\text{total}}}{L}} = 1{,}087.5\ \ \sqrt{\frac{337.7\ \text{fF}}{3\ \text{nH}}} = 11.53$$

If we now assume that all low-frequency upconverted noise is small and further assume that active devices add no noise to the circuit and therefore F =1, we can now estimate the phase noise.

$$\text{PN} = \left[\frac{\omega_o}{(2Q\ \ \omega)} \div \frac{FkT}{2P_S} \div\right]^2$$

$$= \left[\frac{(2\pi 5\ \text{GHz})}{2(11.53)(2\pi 100\ \text{kHz} \div} \frac{(1)(1.38\ \ \ 10^{\ 23}\ \text{J/K})(298\text{K})}{2(67.8\ \mu\text{W})} \div\right]^2 = 1.43 \times 10^{\ 10}$$

This is –98.5 dBc/Hz at a 100-kHz offset, which is 6.5 dB below the promised performance. So, the specifications given to the customer are most likely very difficult (people claiming that anything is impossible are often interrupted by those doing it), given the constraints. This is a prime example of one of the most important principles in engineering. If the sales department is running open loop, then the system is probably unstable and you may be headed for the rails [9].

Example 9.7: Choosing Inductor Size

Big inductors, small inductors, blue inductors, red inductors? What kind is best? Assuming a constant bias current, and noise figure for the amplifier, and further assuming a constant Q for all sizes of inductance, determine the trend for phase noise in a $-G_m$ oscillator relative to inductance size. Assume the inductor is the only loss in the resonator.

Solution:
Since the Q of the inductor is constant regardless of inductor size and it is the only loss in the resonator, then the Q of the resonator will be constant.

The parallel resistance of the resonator will be given by

$$R_p = Q_{\text{ind}}\omega_o L$$

For low values of inductor, R_p will be small. We can assume that the oscillation amplitude is proportional to

$$V_{\text{tank}} \mu R_p$$

We are only interested in trends here, so constants are not important.

Thus, the power in the resonator is given by

$$P_S \mu \frac{V_{\text{tank}}{}^2}{R_P} \mu \frac{(R_p)^2}{R_P} = R_p$$

Now phase noise is

$$\text{PN} = \frac{\omega_o}{(2Q \ \omega)\div}^2 \frac{FkT}{2P_S}\div$$

Q is a constant, and we assume a constant frequency and noise figure. The only thing that changes is the output power.

Thus,

$$\text{PN} \mu \frac{1}{P_S} \mu \frac{1}{L}$$

Thus, as L increases, the phase noise decreases as shown in Figure 9.33.

At some point the inductor will be made so large that increasing it further will no longer make the signal swing any bigger. At this point,

$$P_S \mu \frac{V_{\text{tank}}{}^2}{R_P} \mu \frac{1}{R_P}$$

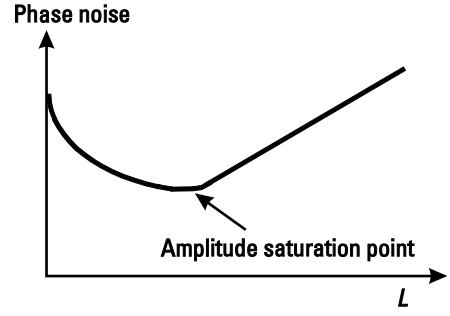

Figure 9.33 Phase noise versus tank inductance.

Again everything else is constant except for the power term, so

$$\text{PN} \propto \frac{1}{P_S} \propto L$$

Thus, once the amplitude has reached its maximum making the inductor any bigger will tend to increase the phase noise. These two curves will intersect at this point. Therefore, we can draw the trend lines as seen in Figure 9.33.

So far, the discussion has been of oscillators that have no tuning scheme. However, most practical designs incorporate some method to change the frequency of the oscillator. In these oscillators, the output frequency is proportional to the voltage on a control terminal:

$$\omega_{\text{osc}} = \omega_o + K_{\text{VCO}} V_{\text{cont}} \tag{9.84}$$

where K_{VCO} is the gain of the VCO (sometimes also called tuning sensitivity) and V_{cont} is the voltage on the control line. If it is assumed that V_{cont} is a low-frequency sine wave of amplitude V_m, and using the narrow band FM approximation, the resulting output voltage is

$$v_{\text{out}}(t) \quad A\cos(\omega_o t) + \frac{A V_m K_{\text{VCO}}}{2 \quad \omega} [\cos(\omega_o + \quad \omega)t \quad \cos(\omega_o \qquad \omega)t] \tag{9.85}$$

where A is the carrier power and ω is the frequency of the controlling signal. Thus, if it is assumed that the sine wave is a noise source, then the noise power present at $\pm \omega$ is given by

$$\text{Noise} = \left(\frac{A V_m K_{\text{VCO}}}{2 \quad \omega} \div \right)^2 \tag{9.86}$$

This can be converted into phase noise by dividing by the signal power:

$$\text{PN} = \left(\frac{V_m K_{\text{VCO}}}{2 \quad \omega} \div \right)^2 \tag{9.87}$$

9.17.2 Some Additional Notes About Low-Frequency Noise

From the preceding analysis, it is easy to see how one might estimate the effect of low-frequency noise on the phase noise of the oscillator. Using a simple small-signal noise analysis, one can find out how much noise is present at the varactor terminals. Then, knowing the K_{VCO}, the amount of phase noise can be estimated.

However, this is not necessarily the whole story. Noise on any terminal, which controls the amplitude of the oscillation, can lead to fluctuations in the amplitude. These fluctuations, if they occur at low frequencies, are just like noise and can actually dominate the noise content in some cases. However, a small-signal analysis will not reveal this.

Example 9.8: Control Line Noise Problems

A VCO designer has designed a VCO to operate between 5.7 and 6.2 GHz, and the tuning voltage is set to give this range as it is tuned between 1.5V and 2.5V. The design has been simulated to have a phase noise of 105 dBc/Hz at a 100-kHz offset. The design has been given to the synthesizer designers who wish to place it in a loop. The loop will have an off-chip RC filter and the tuning line of the VCO will be brought out to a pin. The synthesizer team decides to use a pad with an electrostatic discharge (ESD) strategy that makes use of a 300 series resistor. What is the likely impact of this ESD strategy on this design?

Solution:
First, we estimate the gain of the oscillator:

$$K_{\text{VCO}} = \frac{6.2\text{ GHz} \quad 5.7\text{ GHz}}{2.5\text{V} \quad 1.5\text{V}} = 500\text{ MHz/V}$$

This is a high-gain VCO. As well, it should be noted that this is a very crude estimate of the gain, as the varactors will be very nonlinear. Thus, in some varactor bias regions, the gain could be as much as twice this value.

Next we determine how much noise voltage is produced by this resistor:

$$v_n = \sqrt{4kTr} = \sqrt{4(1.38 \quad 10^{\ 23}\text{J/K})(298\text{K})(300 \quad)} = 2.22\text{ nV}/\sqrt{\text{Hz}}$$

We are concerned with how much noise ends up on the varactor terminals at 100 kHz. Note that it is at 100 kHz, *not* 6 GHz±100 kHz. At this frequency, any varactor is likely to be a pretty good open circuit. Thus, all the noise voltage is applied directly to the varactor terminals and is transformed into phase noise.

$$\text{PN} = \left(\frac{V_m K_{\text{VCO}}}{2 \quad \omega}\right)^2 = \left(\frac{\left(2.22\text{ nV}/\sqrt{\text{Hz}}\right)(500\text{ MHz/volt})}{2(100\text{ kHz})}\right)^2 = 3.08 \quad 10^{\ 11}$$

This is roughly 105.1 dBc/Hz at a 100-kHz offset. Given that originally the VCO had a phase noise of 105 dBc/Hz at a 100-kHz offset and we have now doubled the noise present, the design will lose 3 dB and give a performance of 102 dBc/Hz at a 100-kHz offset. This means that the VCO will no longer meet specifications. This illustrates the importance of keeping the control line noise as low as possible.

9.17.3 Nonlinear Noise

A third type of noise in oscillators is due to the nonlinearity in the transistor mixing noise to other frequencies. For instance, referring to Figure 9.34, assume that there is a noise at some frequency f_n. This noise will get mixed with the oscillation tone f_o to the other sideband at $2f_o - f_n$. This is the only term that falls close to the carrier. The other terms fall out of band and are therefore of much less interest.

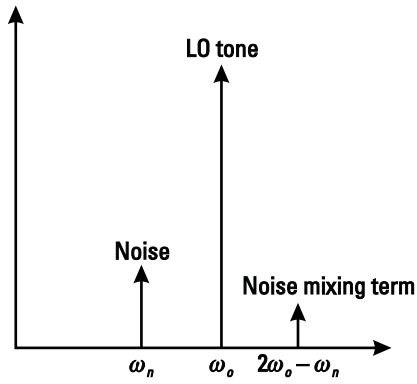

Figure 9.34 Conceptual figure to show the effect of nonlinear mixing.

The magnitude of this noise can be estimated by a power series analysis as in the first edition of this book to show that the two tones are of equal magnitude. Alternatively, one can treat the transistor under large-signal operation as a sampling circuit, and by sampling theory the two tones can also be shown to be of the same amplitude. Thus, Leeson's formula can provide the amplitude of the linear noise, which is present at frequency ω_n. If it is further assumed that there is equal linear noise content at both ω_n and $2\omega_o \quad \omega_n$, then this excess noise is added on top of what was already accounted for in the linear analysis. Since these noise sources are uncorrelated, the powers, rather than voltages, must be added and this means that about 3 dB of noise is added to what is predicted by the linear analysis. Thus, the noise content at an offset frequency is

$$\mathrm{PN} = \frac{A\omega_o}{(2Q \quad \omega)\div}^2 \frac{FkT}{2P_S}\div \tag{9.88}$$

where Q is the quality factor of the resonator, P_s is the power of the oscillator, F is the noise figure of the transistor used in the resonator, k is Boltzmann's constant, T is the temperature of operation, ω_o is the frequency of operation, ω is the frequency offset from the carrier, and A takes into account the nonlinear noise and is approximately $\sqrt{2}$. Note that flicker noise and the thermal noise floor have not been included in this equation but are straightforward to add (see discussion about Figure 9.32).

The term A added to Leeson's formula is usually referred to in the literature as the excess small-signal gain. However, A has been shown, for most operating conditions, to be equal to 3 dB, independent of coefficients in the power series used to describe the nonlinearity, the magnitude of the noise present, the amplitude of oscillation, or the excess small-signal gain in the oscillator.

9.17.4 Impulse Sensitivity Noise Analysis

Because an oscillator is operating with a large signal, the effective transfer function from a noise source to the resulting output phase noise is not constant but is a

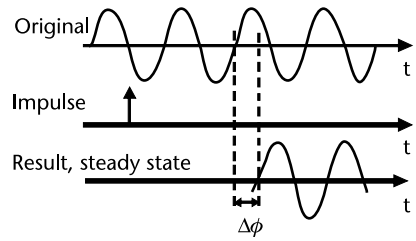

Figure 9.35 The effect of an impulse of noise on the phase of the output.

function of time. For example, noise injected at the zero crossing would produces a large change in phase, and hence a large amount of phase noise while at the peaks, injected noise only produces a small change in phase. This time-varying transfer function, also known as the impulse sensitivity function, is repeated at the rate of the oscillating frequency. In other words, this transfer function of the noise is cyclostationary, similar to the noise in a mixer. The impulse sensitivity function can be determined in a time domain simulation by injecting an impulse at a particular time during the oscillating cycle and determining the resulting steady state shift of phase of the output signal as illustrated in Figure 9.35. This is repeated for impulses injected at various times over the whole period of the oscillating cycle. By combining the noise across the period, the total output noise can be determined. Details are described in [10]. Because this is a fairly time-consuming numerical technique, this is typically not used for quick hand analysis. But it is the basis of various simulation techniques.

Although this technique is more suitable for simulators than for hand analysis, it has been used to derive an analytical expression for noise in Colpitts and $-G_m$ oscillators [11] although the resulting expressions are similar in form to that obtained using Leeson's equation. An interesting result in the paper was that in a CMOS $-G_m$ oscillator, adding a capacitor from the common source node to ground can result in reduced phase noise.

9.18 Making the Oscillator Tunable

Typically, one has the choice of a few different kinds of varactors [12]. The first kind, the pn varactor, is formed from a pn junction (often inside of a well), as shown in Figure 9.36(a). Typically, such varactors have a parasitic varactor to the

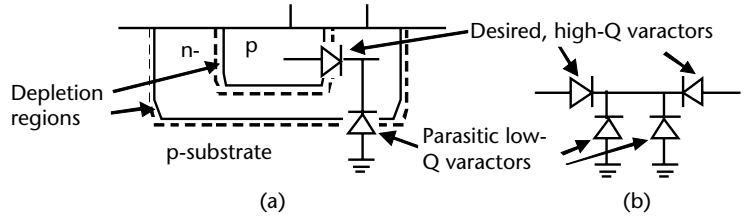

Figure 9.36 (a) Desired pn varactor and parasitic pn varactor, and (b) connecting two pn varactors in such a way that parasitic varactors are at the common-mode point.

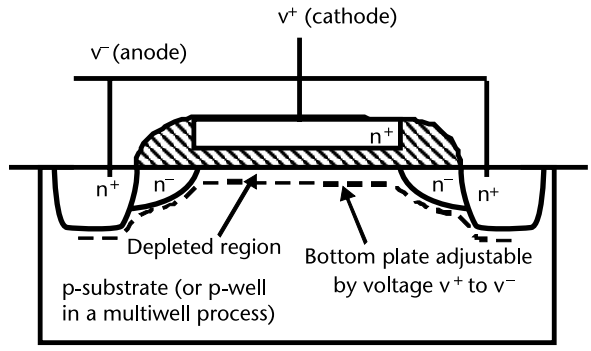

Figure 9.37 Cross section of a MOS varactor in the depletion region.

substrate. Unlike the desired pn junction, which has a high Q, this parasitic junction has a low Q due to the lower doping of the substrate. This makes it desirable to remove it from the circuit. This can be done by placing the varactors in the circuit such that the side with the parasitic diodes is tied together at the axis of symmetry as shown in Figure 9.36(b).

A varactor can also be formed from a regular MOS transistor where the gate is one terminal and the source and drain tied together form the other terminal. Such a structure is shown in Figure 9.37. As a positive voltage is put on the gate, holes in the p- substrate are driven away from the surface, forming a depletion region whose depth depends on applied voltage. At a large enough positive voltage, an inversion layer of electrons forms along the surface and the capacitance is at its maximum value, that of the gate oxide.

If a negative voltage is applied, a layer of holes is accumulated next to the gate oxide. This might be expected to increase the capacitance. However, if the substrate is not connected to the source, there is no direct electrical connection between the accumulation region and the source and drain regions. With a negative gate voltage, there is now a relative positive voltage on the source and drain and as a result, there is a substantial depletion region along the source and drain as shown in Figure 9.38. This depletion capacitance is in series with gate capacitance and as a result, the capacitance does not increase back up to gate capacitance as might have been expected. In this region the low doping in the substrate provides a lossy signal path, hence the Q of such varactors can be quite low.

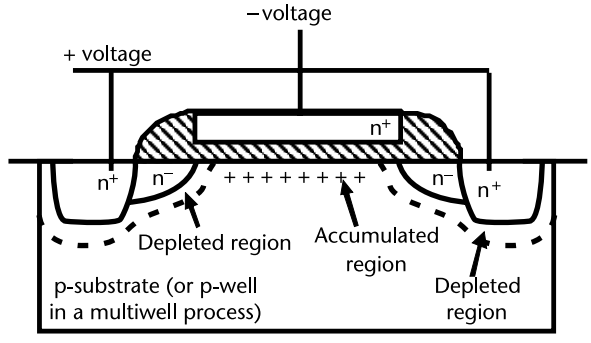

Figure 9.38 Cross section of a MOS varactor shown in the accumulation region.

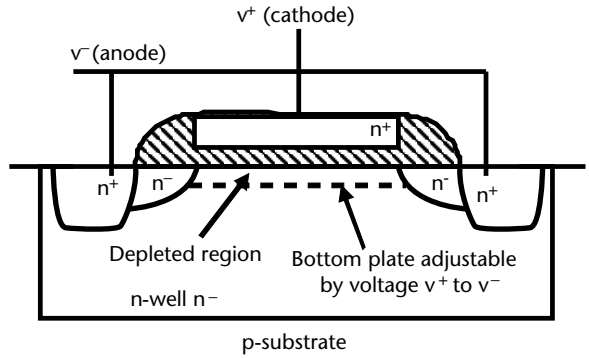

Figure 9.39 Cross section of typical AMOS varactor in the depletion region.

Another technique is to use the gate as one terminal and the substrate connection as the other. In such a case, source and drain are not required. Such a varactor is often called the accumulation MOS, or AMOS varactor, since the capacitance is highest when the surface under the gate is in accumulation. Capacitance is lowest in the inversion region. As with all types of MOS varactors, the AMOS varactor takes advantage of the variation of depletion capacitance with applied voltage. Operation of the AMOS varactor, shown in Figure 9.39, will now be described.

A typical varactor characteristic is shown in Figure 9.40 [13]. The different regions of operation are illustrated in Figure 9.41. When sufficient positive voltage is applied to the gate, a layer of negative charges (electrons) forms along the oxide surface, providing a second electrode to the capacitor and a maximum capacitance equal to the gate capacitance. As the voltage is decreased, this layer of charge disappears and a depletion region forms under the gate oxide. In this region, the depletion width and hence the capacitance is dependent on the gate voltage. For sufficiently large negative voltage on the gate, holes are attracted to the oxide surface and the capacitor is said to be in inversion. In this region, the layer of holes prevents the depletion region from growing any further and the total capacitance is the series

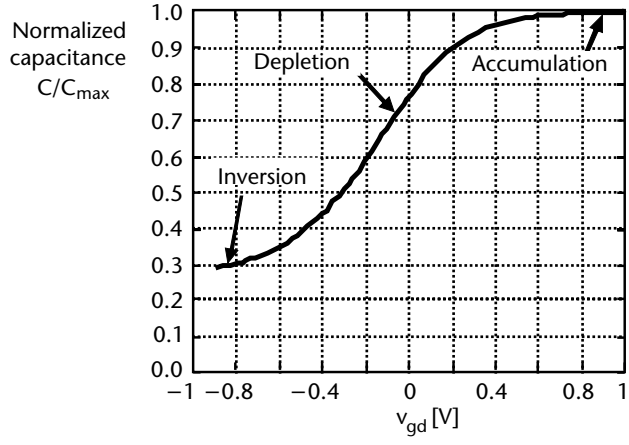

Figure 9.40 Example of a CV curve of an AMOS varactor.

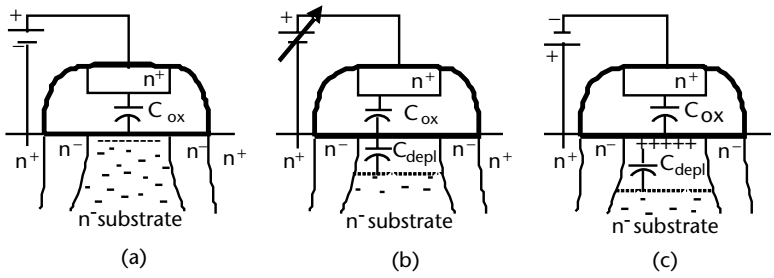

Figure 9.41 Regions of operation in an AMOS varactor (a) accumulation, (b) depletion, and (c) inversion.

capacitance between the gate oxide and the depletion capacitance. Note that while the holes are conductive, they do not connect to the n-type anode and hence do not short out the depletion capacitance.

According to [14], using minimum channel length results in the highest Q varactors, since this minimizes the series resistance. However, with minimum length, the fixed capacitors, such as overlap capacitance, are more important relative to the variable capacitors, so this reduces the ratio of maximum to minimum capacitance. As for Q as a function of region of operation, the series resistance is lowest in the accumulation region where there is a conductive layer close to the gate oxide. As the depletion region is entered, the resistance R will increase; however, in this region the capacitance is decreasing, hence capacitive reactance X_C is increasing. Since Q is the ratio of X_C to R, the increase in R is offset by the increase in X_C, hence Q does not suffer as badly as one might have expected.

Example 9.9: VCO Varactor Placement

Make a differential common-base oscillator tunable. Use base-emitter junctions as varactors and choose an appropriate place to include them in the circuit.

Solution:
The most logical place for the varactors is shown in Figure 9.42. This gives a tuning voltage between power and ground and prevents the parasitic substrate diodes from affecting the circuit.

As can be seen from the last section, low-frequency noise can be very important in the design of VCOs. Thus, the designer should be very careful how the varactors are placed in the circuit. Take, for instance, the G_m oscillator circuit shown in Figure 9.43. Suppose that a low-frequency noise current was injected into the resonator either from the transistors Q_1 and Q_2, or from the current source at the top of the resonator. Note that at low frequencies the inductors behave like short circuits. This current will see an impedance equal to the output impedance of the current source in parallel with the transistor loading (two forward-biased diodes in parallel). This would be given by

$$R_{\text{Load}} = r_{\text{cur}} \,//\, r_{e1} \,//\, r_{e2} \quad \frac{r_e}{2} \tag{9.89}$$

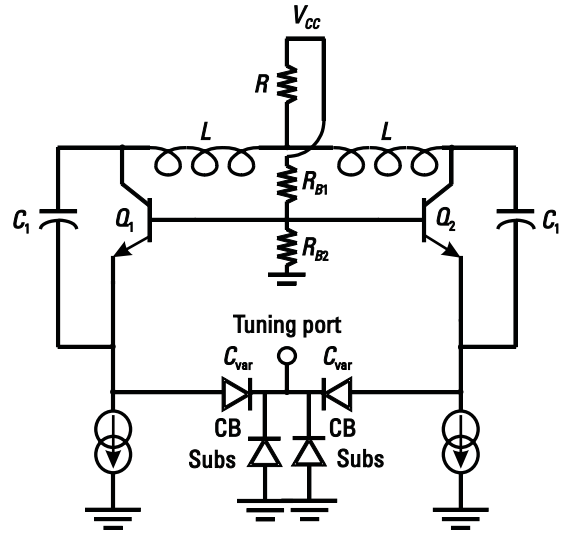

Figure 9.42 A Colpitts common-base design including output buffers and correct varactor placement.

where r_{cur} is the output impedance of the current source and the transistors are assumed to be identical.

This impedance given by (9.89) could easily be in the tens of ohms. Compare this to the circuit shown in Figure 9.23. In the case of Figure 9.23, noise currents have only the dc series resistance of the inductor coil over which to develop a voltage. Thus, if low-frequency noise starts to dominate in the design, the topology of Figure 9.23 would obviously be the better choice.

There are some circumstances where it may not be possible to have inductors connected to a power supply rail. An example is when using a MOS varactor. This varactor

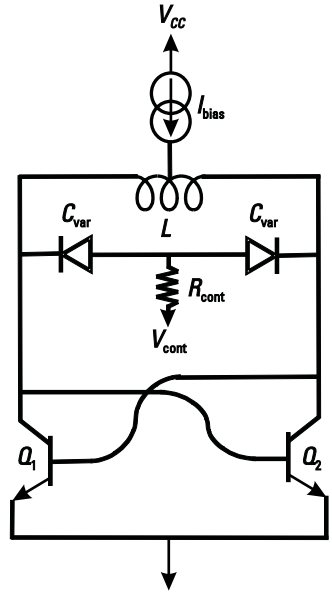

Figure 9.43 An oscillator topology sensitive to low-frequency noise.

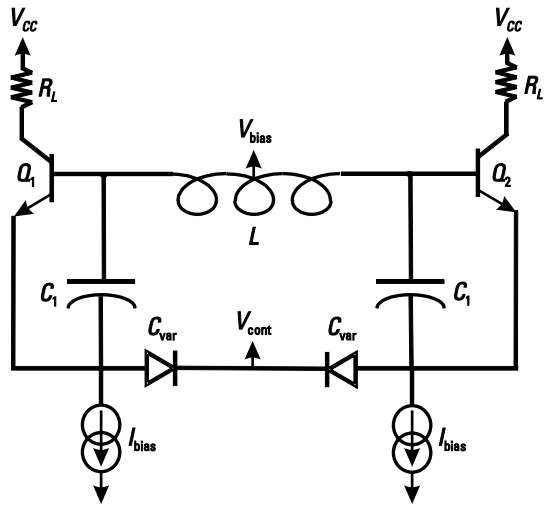

Figure 9.44 A Colpitts common-collector design example.

typically requires both positive and negative voltages. To obtain the maximum tuning range with such a varactor, there is no choice but to place in it in the circuit without the terminals connected to V_{CC}.

Example 9.10: Colpitts VCO Design

Consider the Colpitts VCO shown in Figure 9.44. Design this VCO circuit to meet the specifications given in Table 9.2.

Design

Determining the Capacitor Values to Set the Frequency of Oscillation

The first step is to determine how much capacitance will be needed. We can find the mean total capacitance including parasitics:

$$C_{\text{total}} = \frac{1}{\omega_{\text{osc}}{}^2 L} = \frac{1}{(2\pi \times 2.45\ \text{GHz})^2 4\ \text{nH}} = 1.05\ \text{pF}$$

Table 9.2 VCO Design Specifications

Specification	*Value*
Frequency range	2.4–2.5 GHz
Supply voltage	3.3V
Inductor type	Differential
Inductor value	4 nH
Inductor geometry	Octagonal
Inductor Q at 2.5GHz	14
Varactor type	Base-collector junction
Varactor Q at 2.5 GHz at C_{max} (excludes parasitic diode)	30
Varactor C_{max}:C_{min} ratio	2:1
Load to be driven	50 single-ended or 100 differential

Now the capacitance must be broken between C_1 and C_{var}. Since the r_e of the transistors will load the resonator, we would like to transform this resistance to a higher value. This can be done to greatest advantage if C_1 is made larger than C_{var}. However, with more imbalances in the capacitor sizes, the negative resistance they will generate is lower, and as a result, the oscillator is less efficient. These trade-offs must be carefully considered, preferably with the aid of a simulator. Since we have to start somewhere, let us choose $C_{var} = 0.8C_1$. We can now work out the values for C_1 and C_{var}:

$$2C_{total} = \frac{C_1 C_{var}}{C_1 + C_{var}} = \frac{0.8C_1^2}{1.8C_1} \qquad C_1 = 4.5C_{total} = 4.72 \text{ pF}$$

and therefore $C_{var} = 3.78$ pF nominally, but we still have to work out its min and max values.

$$C_{var\ max} = \frac{2C_1 C_{total\ max}}{C_1 \quad 2C_{total\ max}} = 4.12 \text{ pF}$$

$$C_{var\ min} = \frac{2C_1 C_{total\ min}}{C_1 \quad 2C_{total\ min}} = 3.55 \text{ pF}$$

This is a ratio of only 1.16:1, which means that we can make this capacitor with a smaller varactor in parallel with some fixed capacitance.

This leads to two equations with two unknowns:

$$C_{Fixed} + C_{var\ min} = 3.55 \text{ pF}$$

$$C_{Fixed} + 2C_{var\ min} = 4.12 \text{ pF}$$

solving these two equations, we find that we can replace C_{var} with a 2.98-pF fixed capacitor and a varactor with a C_{max} of 1.14 pF.

Determining the VCO Output Swing

With the frequency of oscillation set and the capacitors sized, the next thing to consider is the large signal behavior of the oscillator. Since the VCO will drive a 50 load for the purposes of this example, thus, the best choice for the R_L resistors would be 50 .

Now we must set the output swing. We would like to make the oscillator swing at the bases of Q_1 and Q_2 as large as possible so that we get good power and therefore help to minimize phase noise. (Note that in a real design there may be many other trade-offs such as power consumption.) The voltage can swing downwards until the current sources start to give out. Most current sources in a bipolar technology (depending on how they are built) will function with a voltage of about 0.5V or more.

Let us choose a swing of about 2 0.8V = 1.6V total.

If we bias the base at 1.65V and it swings down 0.8V from there, then it will sit at about 0.85V. The V_{BE} at this point will be small, < 0.5V (the transistors are off except at the peak of the swing), so the current sources should avoid saturation.

As another check, the emitter swing will only be 0.55·0.8V = 0.44V. So if they are biased at 0.8V, then they should swing between 0.8 and about 0.35 just at the bottom of the swing, which should be fine. The collector is biased at about 100 mV below supply. The peak current in the oscillator will be quite high; the base will swing as high as 2.45V. Thus, the collector can swing as low as 2V without serious problems. The only way that the collector could swing down to 2V is if the current had a peak value of 26 mA or about 15 times the quiescent value. This is not likely.

Setting the VCO Current

Now we need to set the current level to get this voltage swing.
The inductor has a Q of 14 at 2.5 GHz. Therefore, its parallel resistance is

$$R_p = Q_{\text{ind}}\omega_o L = 14\times(2\times\pi\times 2.5\text{ GHz})4\text{ nH} = 880$$

The varactor has a Q of 30 in the $C_{\max}$ condition. This corresponds to a series resistance of

$$R_s = \frac{1}{Q_{\text{var}}\omega_o C_{\max}} = \frac{1}{(2\times\pi\times 2.4\text{ GHz})\times 30\times 1.14\text{ pF}} = 1.94$$

In the middle of the tuning range, the varactors have a capacitance of 0.855 pF. If we assume that the series resistance of these varactors remains constant, then the equivalent parallel resistance is

$$R_{\text{cap}} = \frac{|Z_{\text{var}}|}{R_s}\times\frac{1}{\omega_o C_{\text{var}}}$$

$$= \frac{1}{1.86\quad(2\times\pi\times 2.45\text{ GHz})\times 0.855\text{ pF}}\times\frac{1}{(2\times\pi\times 2.45\text{ GHz})\times 0.855\text{ pF}} = 2.98\text{ k}$$

Now this resistance must be converted into the equivalent resistance at the bases of Q_1 and Q_2 (note that C_2 is made up of 0.855 pF of varactor capacitance and 2.98 pF of fixed capacitance at the middle of the tuning range):

$$R_{\text{cap}} = 2\quad 1+\frac{C_2}{C_1}\div^{2}\quad R_{\text{cap}} = 2\quad 1+\frac{3.83\text{ pF}}{4.72\text{ pF}}\div^{2}\quad 3.1\text{ k}\quad = 20.3\text{ k}$$

Thus, the total parallel resonator resistance is $R_p = 843\quad$.

$$V_{\text{tank}}\quad 2I_{\text{bias}}\quad\frac{C_1}{C_1+C_2}\div R$$

$$I_{\text{bias}} = \frac{V_{\text{tank}}}{2R_{\text{tank}}}\quad\frac{C_1+C_2}{C_1}\div = \frac{1.6\text{V}}{2(843\quad)}\quad\frac{4.72\text{ pF}+3.83\text{ pF}}{4.72\text{ pF}}\div = 1.71\text{ mA}$$

The next thing to do is to size the transistors used in the tank. Transistors were chosen to be 25 μm.

Determining the Phase Noise of the VCO

We can also estimate the phase noise of this oscillator.

The r_e of the transistor at this bias is

$$r_e = \frac{v_T}{I_C} = \frac{25\text{ mV}}{1.71\text{ mA}} = 14.6$$

Since this value will seriously affect our estimate, we also take into account 5 for parasitic emitter resistance (this value can be determined with a dc simulation).

$$R_{\text{re tank}} = 2\ \left(1 + \frac{C_1}{C_2}\right)^2 r_e = 2\ \left(1 + \frac{4.72\text{ pF}}{3.78\text{ pF}}\right)^2 19.6 \qquad 200$$

This can be added to the existing losses to compute the overall resonator resistance of 843 to give 162 .

Now we can compute the Q after increasing C_{total} to account for parasitics

$$Q = R_{\text{tank}}\sqrt{\frac{C_{\text{total}}}{L}} = 162\ \sqrt{\frac{1.18\text{ pF}}{4\text{ nH}}} = 2.78$$

We need to estimate the available power at the tank of the oscillator:

$$P_S = \frac{V_{\text{tank}}{}^2}{2R_{\text{tank}}} = \frac{(1.6\text{V})^2}{2(162\ )} = 7.9\text{ mW}$$

We can now estimate the phase noise of the oscillator. We will assume that the phase noise due to K_{VCO} is not important.

We will assume a noise factor of 1 for the transistor.

$$\text{PN}(f_m) = 10\log\ \left(\frac{Af_o}{(2Qf_m)}\right)^2\ \left(\frac{FkT}{2P_S}\right) = 10\log\ \left(\frac{0.098\text{ Hz}^2}{f_m{}^2}\right)$$

Note that noise voltage is calculated with small-signal resistance values while output voltage is calculated using large-signal resistance. For noise, this is justified since the dominant noise occurs during the zero crossings during which time the transistor nonlinearity is typically not being exercised. A more general approach would consider the noise sensitivity function to quantify the amount of noise being contributed during each part of the output waveform. Note also that in the calculation of phase noise, small-signal resistance is used in calculating the noise—but the amplitude of the carrier was calculated using equivalent large signal resistance. This can be justified by noting that, in reality, on the tank, we are typically concerned with a voltage ratio, rather than a power ratio. Then, later, we convert the tank voltages (both carrier and noise) to an output power using a buffer. Thus, using the small-signal resistance can be seen as turning the noise power back into a noise voltage to allow a direct comparison.

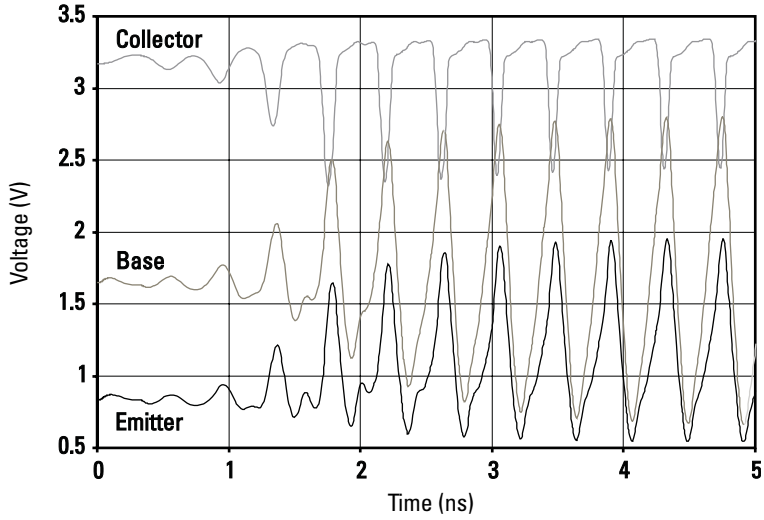

Figure 9.45 Base, collector, and emitter voltages.

VCO Simulation Results

The oscillator was simulated and the following results will now be shown.

Transistor voltage waveforms can be seen in Figure 9.45. Note that the transistor safely avoids saturating. Also note that emitter voltage stays above 0.5V, leaving the current source enough room to operate.

The collector current waveform is shown in Figure 9.46. Note how nonlinear the waveforms are. The transistor is on for only a short portion of the cycle.

The differential collector output voltage is shown in Figure 9.47. Notice all the harmonics present in this waveform. After looking at the collector currents, this is to be expected.

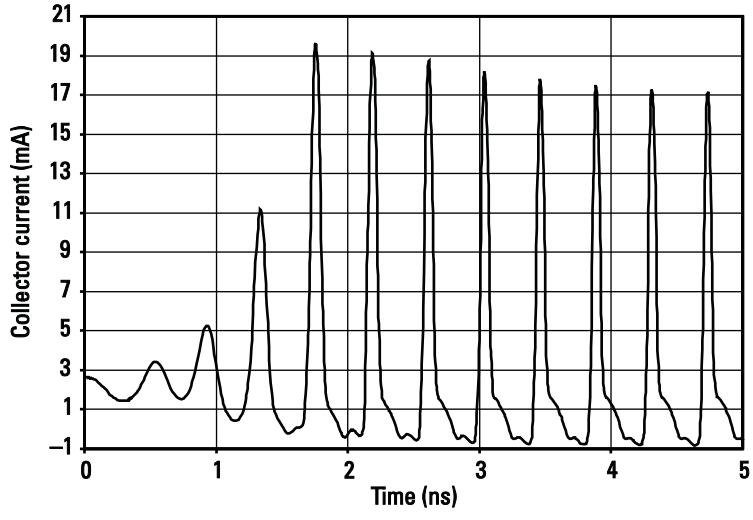

Figure 9.46 Collector current.

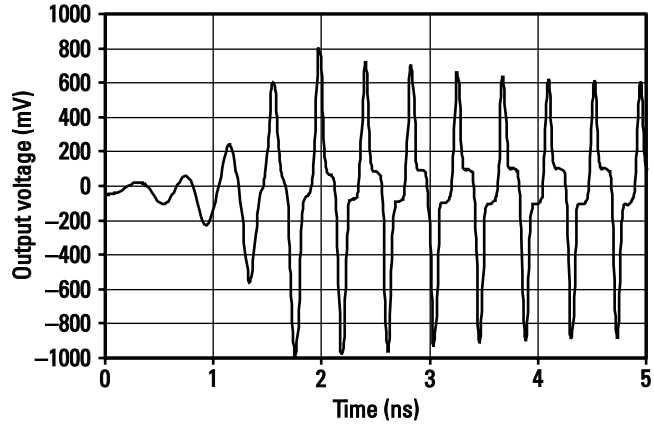

Figure 9.47 Differential output voltage waveform.

The differential resonator voltage is shown in Figure 9.48. It is much more sinusoidal as it is filtered by the LC resonator. The peak voltage is around 1.8V, which is only slightly higher than the simple estimate (1.6V) we did earlier.

The tuning voltage versus frequency for the VCO is shown in Figure 9.49. Note that we are only getting about 60 MHz of tuning range and that the frequency has dropped about 120 MHz compared to the design calculations. This clearly shows the effect of transistor parasitics. This design will have to be tweaked in simulation in order to make its frequency accurately match what was asked for at the start. This is best done with a simulator, carefully taking into account the effect of all stray capacitance. The results of the phase noise simulation and calculation are shown in Figure 9.50.

Example 9.11: CMOS VCO Design

Design a $-G_m$ VCO in a commercial 130-nm CMOS process to operate from 7.0 to 7.3 GHz with an oscillation amplitude greater than a 400-mV peak differential

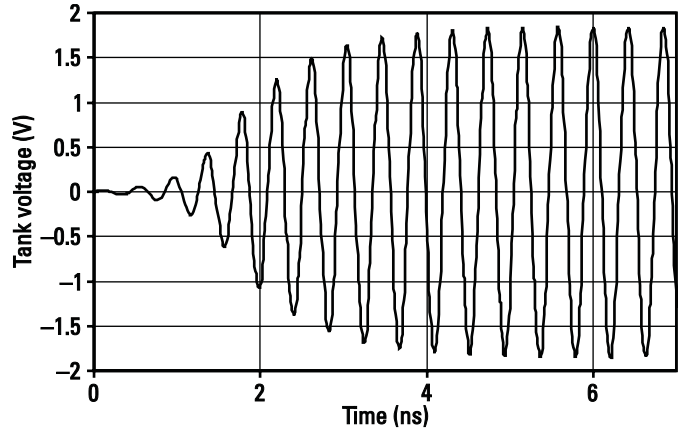

Figure 9.48 Differential resonator voltage.

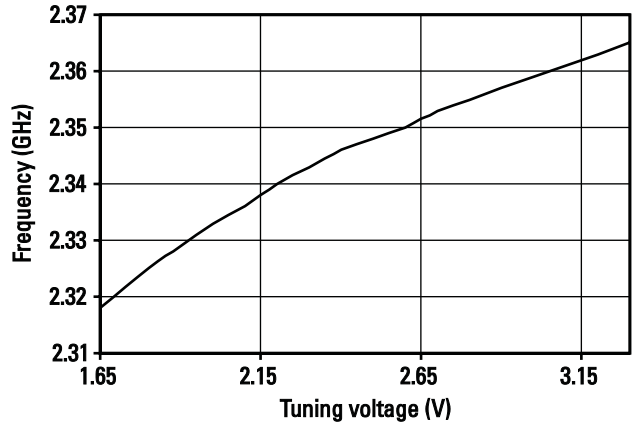

Figure 9.49 Frequency versus varactor tuning voltage.

at the output and a phase noise of better than 110 dBc/Hz at a 1-MHz offset. Assume that $R_L = 500$.

Solution:
The oscillator schematic is shown in Figure 9.51.

Bias Currents and Transistor Sizes to Achieve Required Voltage Swing
The first step is typically to set the bias current and the transistor size to achieve the desired amplitude. Since the buffer will have significant loss we will aim for V_{tank} of about a 700-mV peak differential so that including losses the output amplitude will be greater than the required 400 mV, and if necessary we can adjust it later. Using an estimated inductor Q of 15 and assuming the varactor Q will be significantly higher, the bias current is calculated as

$$v_{\text{tank}} = 0.7V = \frac{2}{\pi} I_B R_T = \frac{2}{\pi} I_B Q_T \omega L \qquad I_B L = 1.67 \quad 10^{-12}$$

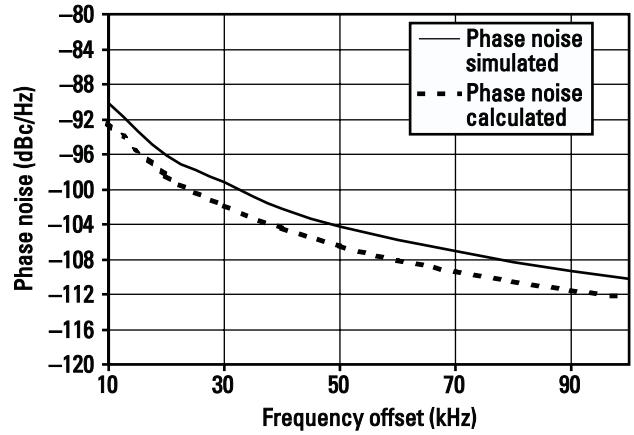

Figure 9.50 Simulated and calculated phase noise.

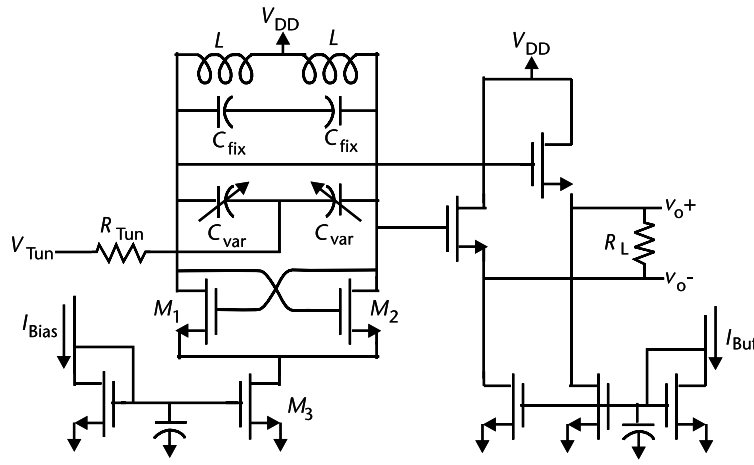

Figure 9.51 CMOS oscillator schematic.

This provides the required current, inductance product. To save power, low current and high inductance is picked. However, the maximum inductor size is limited by the need to keep the self-resonant frequency well above the operating frequency, and with low currents it is important to make sure that the transistor g_m is large enough for oscillations to start. We choose a starting point of 1 mA or 0.5 mA/side and 1.67 nH total, or 835 pH/side after it was determined that the inductor is still feasible at 7 GHz. With 1.67 nH and an assumed Q of 15, the inductor total parallel resistor would be 1,102 at 7 GHz and 1,149 at 7.3 GHz, to be updated later with simulations including the varactors.

The required transconductance is given by

$$g_m > \frac{2}{R_T} = 1.8 \text{ mA/V}$$

Using a minimum length transistor with a width of 20 μm, the same transistor characterized in Chapter 4, it was determined that at 0.5 mA, the transconductance is about 8 mA/V (Figure 4.20); thus, it is well over the minimum required to oscillate. We also note from the simulated transistor f_T of about 50 GHz (Figure 4.21) and the g_m that the parasitic capacitance is about 26 fF.

Setting Frequency and Tuning Range

Now we must choose capacitors and varactors to achieve the correct frequency and tuning range. Capacitance and frequency are related by

$$f_o = \frac{1}{\sqrt{LC}}$$

To achieve the required range from 7 to 7.3 GHz, we can determine the required capacitance ratio as

$$\frac{C_{\max}}{C_{\min}} = \frac{C + C_{\text{var,max}}}{C + C_{\text{var,min}}} = \left(\frac{f_{\max}}{f_{\min}}\right)^2 = \left(\frac{7.3}{7.0}\right)^2 = 1.09$$

From varactor simulations, it was determined that the AMOS varactors available in this process could achieve a Q of about 30 or over if measured from the gate side, and a C_{max}/C_{min} ratio of about 2.5 if biased at 0.6V; however, this ratio drops to only about 1.4 if the nominal bias voltage is 1.2V, as it would be for the simplest NMOS $-G_m$ oscillator. Since we only need a ratio of 1.09 this is adequate, in fact there is room to use a fixed high-Q MIM capacitor in parallel with the varactors to increase the loaded tank Q and to lower the oscillator gain, both of which would help to improve phase noise. The required capacitances are determined to be 619 fF/side at 7 GHz and 569 fF at 7.3 GHz. To use the full varactor range, we can use a varactor with a minimum value of $C_{var,min}$ = 125 fF, and a maximum value of $C_{var,max}$ = 175 fF, and place this varactor in parallel with a fixed capacitor of value C = 444 fF. To be safe we can choose a somewhat larger varactor, for example, from about 190 to 270 fF; hence, C = 364 fF. Note that C will be adjusted to compensate for parasitic capacitances. With a Q of 30, the varactor has an equivalent parallel resistance of 2.53 k /side at 7 GHz and 3.44 k /side at 7.3 GHz. Combining varactor and inductor parallel resistors results in a total parallel resistance of 905 at 7 GHz and 985 at 7.3 GHz and the effective Q is reduced to 12.32 and 12.86 for 7 and 7.3 GHz, respectively.

Estimating Phase Noise and Scaling Transistor Sizes to Meet the Noise Specification

The next step is to estimate the phase noise and scale the transistors as necessary to meet the specifications. The equation for phase noise is

$$\text{PN} = \frac{A\omega_o}{(2Q \quad \omega)}\div \quad \frac{FkT}{2P_S}\div$$

At 7 GHz, $A = \sqrt{2}$, $\omega_o = 2\pi$ 7 GHz, $Q = 12$, $\omega = 2\pi$ 1 MHz, and P_s is given by

$$P_S = \frac{v_{\text{tank}}^2}{2R_p} = \frac{0.7^2}{2\times 905} = 271\ \mu\text{W}$$

And assuming that $1/f$ noise is not dominant and the current source is designed properly, F is estimated by

$$F = 1 + 4\gamma g_m R_p(1 \quad \rho) = 1 + 4\times 0.67\times 8\text{m}\times 905\times 0.2 = 4.88$$

where the long channel value for γ has been used since the oscillator is operating at low current density and (1 ρ), the transition time as a fraction of the period has been estimated as 20%. It should be apparent that if the transition time is faster (e.g., 10%), the noise factor will be substantially reduced. Solving for the phase noise, using k = 1.38 10 ^{23} and T = 298K results in:

$$\text{PN} = \frac{\left(\sqrt{2}\times 7\text{ GHz}\right)}{2(12.32)(1\text{ MHz})}\div \quad \frac{(4.88)(1.38 \quad 10^{\ 23}\text{ J/K})(298\text{K})}{2(271\ \mu\text{W})}\div$$

$$= 5.977\times 10^{\ 12} \qquad 112\text{ dBc/Hz}$$

This is adequate, but with a number of assumptions made. In addition, there is the phase noise due to the tuning line. For this, the oscillator gain must be known. From the varactor characterization, at the lowest frequency when the capacitance is largest, the capacitance is nearly independent of voltage; hence, phase noise implications will be minimal. At the other end the slope is maximum when the capacitance is smallest and it is estimated that the VCO gain is about 1.2 GHz/V. Using this value, and a tuning resistor of 1 k , the resulting phase noise is

$$\text{PN} = \frac{V_m K_{\text{VCO}}}{2\ \omega}\div^{2} = \frac{4.06\text{n}\times 1.2G}{2\times 1\text{M}}\div^{2} = 5.93\times 10^{-12} \qquad -112 \text{ dBc/Hz}$$

This is roughly equal to the previously calculated noise, hence it is expected that phase noise will rise by about 3 dB at 7.3 GHz compared to 7.0 GHz.

Preliminary Simulations and Further Transistor Sizing for Noise

At this stage preliminary simulations were done with Spectre and with all transistors at minimum length and a width of 20 μm, it was discovered that 1/*f* noise was by far the dominant noise source, hence the oscillator was not anywhere near the specification (recall the above calculations were done assuming 1/*f* noise was not dominant). The main contributors were the bias transistors with some contribution from the rest. To reduce 1/*f* noise, transistor gate area must be increased; hence, the bias transistor sizes were increased to 200/1 μm. Other transistors were left at minimum length, but the widths of the cross-coupled transistors were increased to 30 μm. The current does not change since the circuit is being driven with a current source. It was noted that while the tank voltage was close to the desired 700-mV differential, the output voltage was low due to the low transconductance of the buffer transistor. This transistor width was increased to 50 μm and its bias current was increased to 2 mA to result in an output swing of over 400 mV as required by the specification. On each side of the resonant circuit there are parasitic capacitance from the cross-coupled transistors at $C_{GS} + 4C_{GD}$ of about 66 fF/side (estimated C_{gs} of 30 fF and C_{gd} of 9 fF) and the buffer with estimated C_{gs} of 50 fF for a total of about 126 fF. Thus, the original value of 364 fF is reduced to 238 fF/side. By simulation, the capacitors needed to be increased to 300 fF/side because the actual inductor turned out to be only 1.5 nH rather than the originally planned 1.67 nH. But because the varactors had originally been made somewhat bigger than the exact minimum value, it was still possible to cover the required tuning range.

Note that in the interest of saving power, the transistors are operated at very low current density and the oscillator is current limited. At such low current, neither the transistors nor the oscillator are at their lowest noise, so if lower noise were required it would be possible to increase the current. This would have the effect of raising the output signal levels more than the output noise; hence, phase noise would be reduced. It is also possible to increase g_m without changing the bias current by increasing the transistor width. This results in a larger noise source, but can increase switching speed since a smaller v_{GS} is required to switch (estimate linear region in a differential pair as I_o/g_m). Faster switching results in smaller value for (1 ρ).

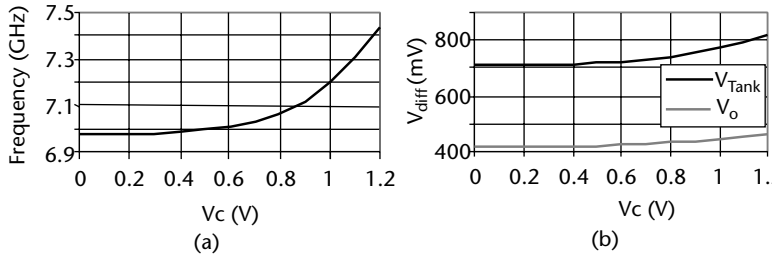

Figure 9.52 Oscillator performance (a) frequency and tuning range and (b) voltage swing v_{Tank} and v_{out}.

Computer Simulation Results

Simulations were then done to verify performance. First it was verified that the oscillator could operate across the desired frequency range. Figure 9.52(a) shows frequency from about 6.98 GHz to about 7.43 GHz. It also verifies that the slope is about 0 at 7 GHz while it is about 1.2 GHz/V at 7.3 GHz. Figure 9.52(b) shows that the resonator voltage v_{Tank} is more than 700 mV and the output voltage v_{out} is more than 400 mV across the whole tuning range.

Then phase noise is plotted at 7 GHz as a function of offset frequency in Figure 9.53. At a 1-MHz offset, the phase noise is about 111 dBc/Hz. By examining the intersection of the –30-dB/decade slope with the 20-dB/decade slope the 1/*f* corner is seen to be at about 100 kHz.

Figure 9.54(a) shows the phase noise as a function of control voltage, and hence of oscillator frequency. As predicted, the noise rises by a little over 3 dB by 7.3 GHz. Thus it is no longer meeting the specification. However, by reducing the tuning resistor, this component of noise should be greatly reduced. For example, by reducing the resistor by a factor of 4 times to 250 , this component of noise should be down by 6 dB, and it would likely meet the specification of 110 dBc/Hz. Figure 9.54(b) shows the currents in the cross-coupled transistors showing that they are behaving approximately as switches. As well, it can be seen that there is a second harmonic component on the total current and it is not constant at 1 mA as designed

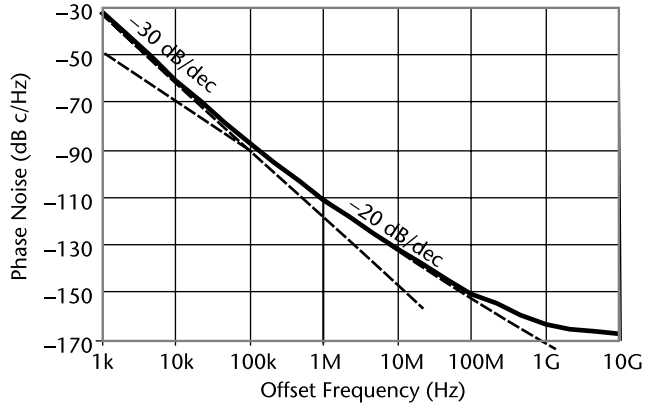

Figure 9.53 Phase noise as a function of offset frequency for a 7-GHz oscillator.

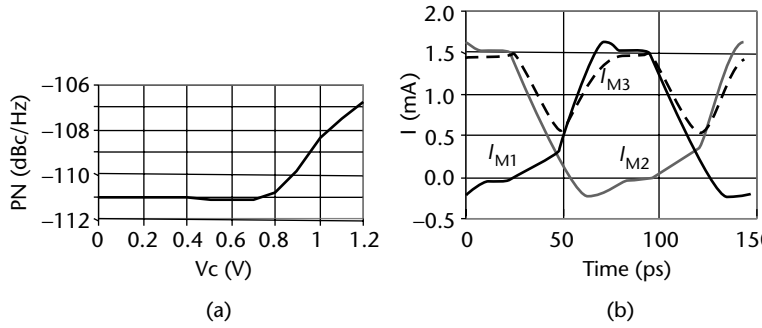

Figure 9.54 (a) Phase noise at a 1-MHz offset as a function of control voltage and (b) currents in cross-coupled transistors M_1 and M_2 and in current source M_3.

but is swinging from about 0.5 to 1.5 mA with an average value of about 1.1 mA. The transition time for each transistor can be estimated as about 10% of the total period, thus the noise figure prediction could be modified accordingly—the oscillator noise factor would be decreased.

As well, a final noise summary was done to compare to the predicted noise factor. Considering that the source noise is due to the parallel resistances of the tank inductors and varactors, the noise factor was simulated to be 2.92. This is relatively close to the estimate taking into account the faster transition time. However, in the summary it is clear that there are many components to the noise and the drain channel noise is about a third of the total noise. Nearly a third comes from 1/*f* noise and the final third is a combination of all other noise sources including the buffer transistors. Thus our simplistic estimate for noise will never be exact, but it does successfully predict approximate performance.

9.19 Low-Frequency Phase-Noise Upconversion Reduction Techniques

The following three sections discuss three techniques to reduce phase noise up conversion. These will consist of bank switching, differential tuning, and simultaneous matching of transconductance and impedance.

9.19.1 Bank Switching

To cover the required frequency band and the effects of process variations, a particular K_{VCO} is required. However, if the band is broken into many subbands, then K_{VCO} is reduced and phase noise upconversion is reduced. For example, by breaking a band into three subbands, potentially K_{VCO} is reduced by a factor of 3 and phase noise upconversion could be reduced by a factor of about 10 dB. One of the better ways to implement bank switching with n banks of varactors as shown in Figure 9.55 is to use (n 1) AMOS varactors as switches and one bank as a continuously variable capacitor. Because of the shape of the AMOS varactor curve (as shown in Figure 9.40) when operated as a switch, AMOS varactors have a low gain when fully switched. This will result in minimal low-frequency noise

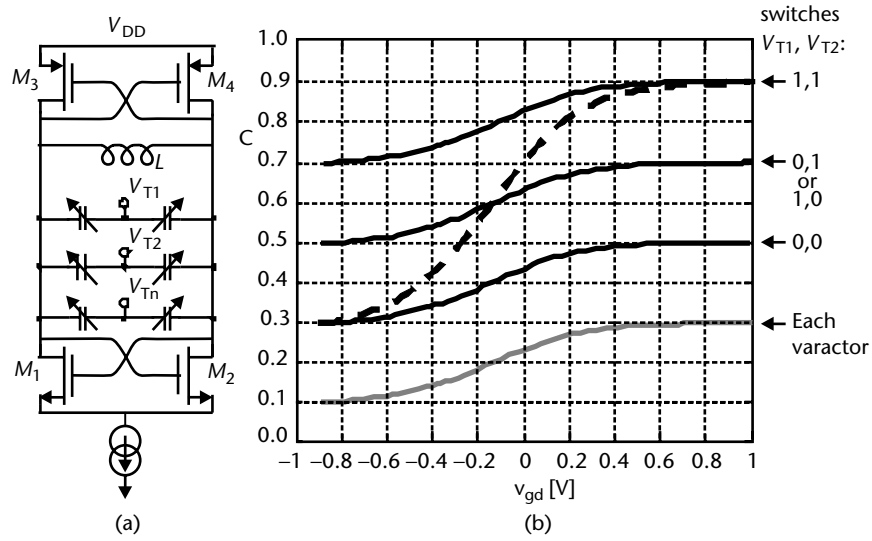

Figure 9.55 Oscillator with banks of varactors. To minimize low-frequency noise upconversion, the recommended operation is to have n 1 operated as switches and one for continuous tuning.

upconversion. Note that to get full switching, both positive and negative control voltages must be supplied. This is conveniently obtained when the oscillator has both NMOS and PMOS cross coupling, since the output nodes are nominally biased between the power supply voltages. However, for the case where there is only a single polarity of cross-coupling, the output voltage will be nominally at the rail voltage (e.g., with NMOS cross coupling as in Figure 9.21(c), the outputs are nominally at V_{DD}). In such a case, the varactors will need to be decoupled in order to obtain both positive and negative bias voltages. Such additional capacitance in series will reduce the K_{VCO}, which can be accounted for in the design and is not a problem, as lower gain was desired; however, it will limit the total tuning range.

It is possible to use weighted varactor sizes to get more curves with a smaller number of capacitors and control voltages. An example is shown in Figure 9.56 with two switchable varactors (with capacitance values of 50/150 fF, 150/450 fF). As well, there is one adjustable varactor, adjustable between 100 and 300 fF. As shown, there are four possible output curves with each varying by 100 fF, with a total range of capacitance values between 300 and 900 fF, the same range as before; however, there is now some overlap between the ranges, allowing for a more flexible oscillator design.

An automated method can be devised to tune such a circuit within a PLL-based synthesizer. A comparator-based circuit is used to observe the VCO control voltage. If the control voltage approaches one of its limits, a comparator output triggers the next curve in the appropriate direction to be selected, for example, by either counting up or counting down in a binary counter. The counter output can then select the appropriate switchable varactors. Note that binary weighting works well if varactors have capacitance ratio of 2:1. In such a case, it is straightforward to get uniform curves. The example above illustrates

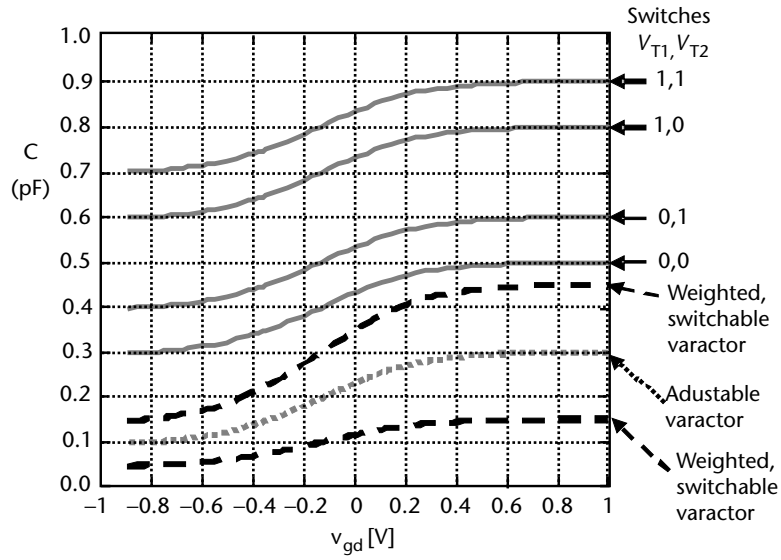

Figure 9.56 Oscillator with weighted switched varactors and an adjustable varactor.

that with a 3:1 capacitance ratio, there is some difficulty in getting uniform curves.

9.19.2 g_m Matching and Waveform Symmetry

It is often stated that phase noise upconversion can be reduced by matching the transconductance of PMOS and NMOS negative g_m cells in a complementary-style VCO such as the one shown in Figure 9.21(e). This reduction occurs since any disturbance at the output will produce equal currents in the PMOS and NMOS transistor, thus minimizing the effect on the output. However, this is not the full story, as it is possibly more realistic to consider a noise current being injected into the output, for example, because of $1/f$ noise in one of the transistors. This current injected into the output node is completely canceled out if both the g_m and the impedance are matched. Typically, to match g_m requires larger PMOS transistors than NMOS transistors; hence, the gate capacitance is larger and the impedance will be lower. However, if the process is fast enough that one can use a nonminimum gate length in the NMOS transistors, then a simultaneous match is possible. For example, in [15] it was demonstrated that simultaneous g_m and capacitance matching resulted in up to 8 dB of phase noise improvement compared to matching for g_m only. Mathematically, matching both the capacitance of both NMOS and PMOS transistors in the design requires that:

$$\begin{aligned} \gamma W_n L_n C_{ox} &= \gamma W_p L_p C_{ox} \\ W_n L_n &= W_p L_p \end{aligned} \tag{9.90}$$

where W_n and L_n are the width and length of the NMOS transistors and W_p and L_p are the width and length of the PMOS transistors. If the g_m of both NMOS and PMOS are to be the same, it requires that:

$$\sqrt{2\mu_n C_{\text{ox}} \frac{W_n}{L_n} I_{\text{DS}}} = \sqrt{2\mu_p C_{\text{ox}} \frac{W_p}{L_p} I_{\text{DS}}}$$
$$\mu_n \frac{W_n}{L_n} = \mu_p \frac{W_p}{L_p} \tag{9.91}$$

These two conditions mean that for best phase noise, the ratio of the lengths of the transistors should be:

$$L_n = L_p \sqrt{\frac{\mu_n}{\mu_p}} \tag{9.92}$$

where L_p is usually set to the minimum allowed by the technology and the ratio of the width of the PMOS and NMOS transistors should be:

$$\frac{W_p}{W_n} = \sqrt{\frac{\mu_n}{\mu_p}} \tag{9.93}$$

9.19.3 Differential Varactors and Differential Tuning

Differential varactors controlled by a differential tuning voltage as shown in Figure 9.57(a) can be used to reduce or eliminate low-frequency noise upconversion. This is shown in Figure 9.57(b), which shows varactor capacitance versus tuning voltage. The varactors labeled $C_{\text{var+}}$ have capacitance that increases with applied voltage, while the varactors labeled C_{var} decrease with applied voltage. Thus, if a differential voltage is applied, $C_{\text{var+}}$ diodes see a positive voltage, while C_{var}

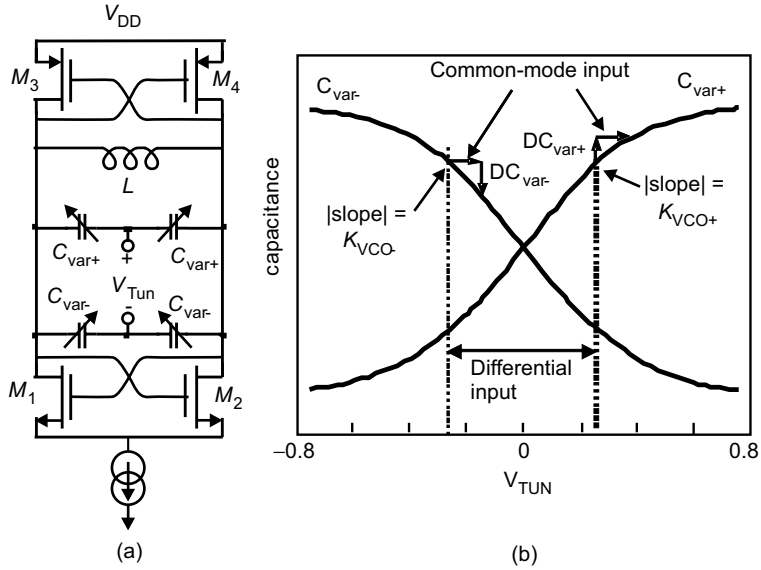

Figure 9.57 Oscillator with differential varactor tuning: (a) schematic and (b) capacitance versus tuning voltage.

varactors see a negative voltage. Hence, both varactors are increased in capacitance for an increase in differential input voltage. However, for a common-mode voltage, both varactors see a voltage in the same direction; hence, the increase in capacitance from the C_{var+} varactor is matched by an equal decrease in the capacitance from the C_{var} varactors. Low-frequency noise (e.g., $1/f$ noise injected from the cross-coupled transistors or the bias circuit) is equivalent to a common-mode input due to the low impedance of the inductor at low frequencies. With a positive varactor slope given by K_{VCO+} and negative varactor slope given by K_{VCO} , low-frequency noise upconversion is given by:

$$\text{PN} = \frac{V_m(K_{VCO+} \quad K_{VCO})}{2 \quad \omega}^{2} \div \tag{9.94}$$

Thus, common-mode or low-frequency noise rejection is only effective if the differential varactors are exactly symmetrical. With perfect symmetry, there is complete rejection; however, for any residual error in symmetry, rejection is reduced. To optimize symmetry, a nonzero bias voltage may be required on the varactor. For example, directly using the varactor characteristics as shown in Figures 9.55 and 9.56 would not produce a symmetrical result, since at zero volts, the capacitance is not halfway between the minimum and maximum capacitances. However, using a dc bias of 0.15V results in operation at the point of average capacitance and maximum symmetry, and this is how the curves in Figure 9.57(b) were generated. Nonsymmetry can be the result of the difference between the parasitic capacitance on the two ends of the varactor. As an example, in [14], 5-GHz oscillators using differential AMOS varactors showed a phase noise reduction of about 10 dB compared to oscillators using single-ended AMOS varactors. The varactors were in an SOI process that was shown to have better symmetry than bulk CMOS; hence, differential tuning in bulk CMOS would be expected to show a somewhat lower phase noise improvement.

Example 9.12: PMOS Versus Complementary VCO Design

Compare the design of PMOS only to a complementary VCO topology as in Figure 9.21(d) and Figure 9.21(e). Design each for the best swing to achieve the lowest phase noise. The VCOs should oscillate at 5 GHz in a 0.18-μm CMOS technology. A 1-nH inductor with a Q of 10 is available for the design. Assume for this design that the mobility of the PMOS transistors is one-third that of the NMOS transistors and design the VCOs to be powered off a 1.5-V supply.

Solution:

A 1-nH inductor with a Q of 10 will have 314 of parallel resistance. Getting the VCO to oscillate at 5 GHz with a 1-nH inductor will mean that the total capacitance due to the transistors plus additional capacitance should be about 1 pF. Now the VCO current must be set appropriately. With the complementary design, the voltage at the resonator will swing about mid-rail to the supply and down to ground, leaving some room for the current source to operate correctly. If the current source can be operated with 300 mV of headroom, this will mean that the

maximum peak-to-peak swing for this design will be 1.2V. The current should be set to give this amplitude for lowest phase noise, but more current will be wasted and lead to excess noise. Therefore, for the complementary design the current should be set so that the peak voltage swing is 0.6V. Therefore:

$$I_{\text{comp}} = \frac{2R_P I}{\pi \times v_{\text{out}}|_{\text{Comp}}} = 3 \text{ mA}$$

In the case of the PMOS design, leaving 300 mV for the current source, the peak swing should be 1.2V. Therefore, the current should be set to:

$$I_{\text{PMOS}}\, v_{\text{out}}|_{SE} = \frac{R_P I}{\pi \times v_{\text{out}}|_{\text{PMOS}}} = 12 \text{ mA}$$

Now with twice the swing, but exactly the same resistance and Q, that means that theoretically the phase noise of the PMOS design should be 6 dB better than the phase noise of the complementary design, but at the cost of four times the power. If the PMOS design were chosen to have the same swing as the complementary design, then it would draw 6 mA of current, thus delivering the same phase noise as the complementary design, but at twice the power.

In order to build these designs, the transistors must be sized. For the complementary design, the PMOS transistors are chosen to have a length of 0.18 μm and the NMOS transistors are chosen to have a length of:

$$L_n = L_p \sqrt{\frac{\mu_n}{\mu_p}} = 0.31 \ \mu\text{m}$$

The widths of the transistors must be set large enough so that the $1/f$ noise is not too large, and so that the on resistance of the devices does not act to limit the swing excessively, but not so large that the parasitic capacitance of the devices dominates the frequency of oscillation. In this design, the width of the NMOS was chosen to be 55 μm, while the PMOS width was set to be 110 μm (note that this is a ratio of 2:1, which is different than the theory in order to get both the C_{gs} and the g_m to match closely in simulation). In the case of the PMOS-only design, the transistors were made wider (175 μm) in order to accommodate the larger currents. The single-ended output voltage for each of the designs is shown in Figure 9.58. This shows that the designs have close to the predicted output swings. The PMOS design with 12 mA is a little lower due to the finite source-drain resistance of the PMOS at higher currents.

The phase noise for the three designs is shown in Figure 9.59. As predicted, the phase noise of the PMOS VCO and the complementary VCO are identical when they have the same output swing, although the PMOS VCO burns twice as much current. When the PMOS design is run at full swing, it delivers 4 5 dB better phase noise then the complementary design is capable of delivering. Note that this is slightly less than the 6 dB that the simple theory would predict because more current means that more noise is produced in this design.

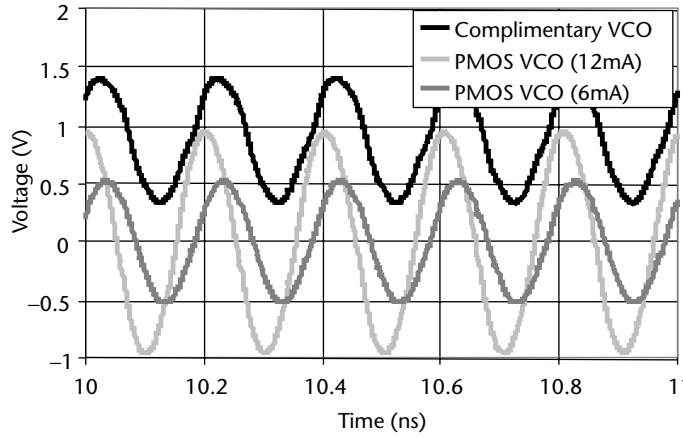

Figure 9.58 Comparison of VCO output voltage.

9.20 VCO Automatic-Amplitude Control Circuits

The purpose of adding automatic-amplitude control (AAC) to a VCO design is to create a VCO with good phase noise and very robust performance over process, temperature, and frequency variations [16]. A VCO schematic with simple additional feedback circuitry is shown in Figure 9.60.

The current in the VCO is set by transistor Q_6, which acts as a current source. Transistors Q_3 and Q_4 are used to limit the swing of the oscillator to $\pm 2\ V_{BE}$. Once the oscillation gets this large, these transistors start to turn on briefly at the top and bottom of the oscillator's swing, loading the resonator with their dynamic emitter resistance. This will effectively de-Q the circuit and prevent the signal from growing any larger. This will prevent transistors Q_1 and Q_2 from entering staturation. However, if transistors Q_3 and Q_4 have to be heavily turned on to limit the swing, they will also start to affect the phase noise performance of the circuit.

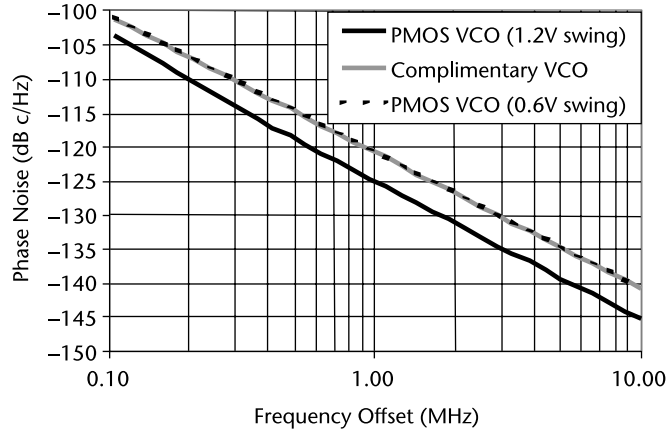

Figure 9.59 Comparison of VCO phase noise.

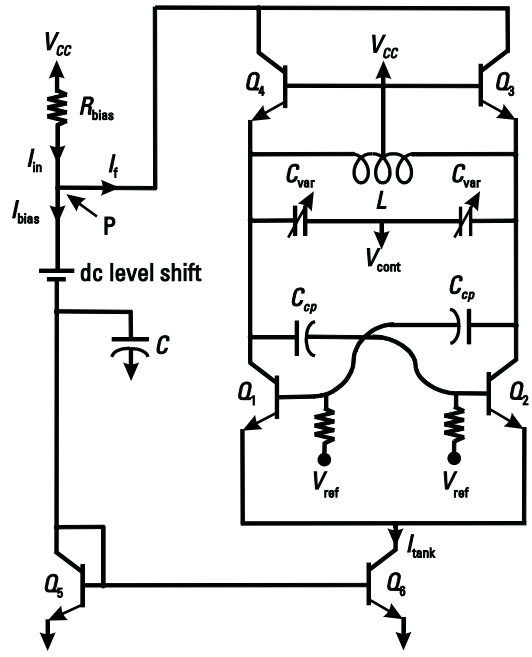

Figure 9.60 VCO topology with feedback in the bias to control the amplitude.

Transistors Q_3 and Q_4 limit the amplitude of the oscillation directly, but are also the basis for the feedback loop that is the second mechanism used to make sure that the VCO is operating at an optimal level. Once these transistors start to turn on, they start to draw current I_f. Their collectors are connected back to the resistor R_{bias}. Q_3 and Q_4 then steal current away from the bias, causing the current in Q_6 to be reduced. This in turn reduces the current in the VCO. Since the VCO amplitude is related to its current, the amplitude of the VCO is thus reduced until the transistors Q_3 and Q_4 just barely turn on. This ensures that the VCO always draws just enough current to turn on these transistors and no more, although the reference current through R_{bias} may vary for any number of reasons. The reference current through R_{bias} must therefore be set higher than the optimum, as the loop can only work to reduce the current through the oscillator, but can never make it higher. Put another way, unless the transistors Q_3 and Q_4 turn on, the loop has zero gain.

The loop can be drawn conceptually as shown in Figure 9.61. The point P shown in Figure 9.60 acts as a summing node for the three currents I_{in}, I_{bias}, and I_f. The current mirror amplifies this current and produces the resonator current, which is taken by the VCO core and produces an output voltage proportional to the input resonator current. The limiting transistors at the top of the resonator take the VCO amplitude and convert it into a current that is fed back to the input of the loop.

The transfer function for the various blocks around the loop can now be derived. In the current mirror, a capacitor C has been placed in the circuit to limit the frequency response of the circuit. It creates a dominant pole in the system (this helps to control the effect of parasitics poles on the system) and limits the frequency response of the loop. I_{bias} generates a voltage across the parallel combination of the

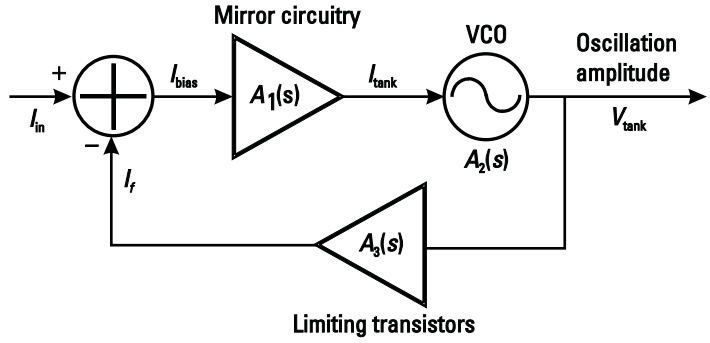

Figure 9.61 Conceptual drawing of the AAC feedback loop.

diode connected transistor Q_5 and C. This voltage is converted into the current I_{tank} by Q_6. Thus, the transfer function for this part of the loop is given by

$$A_1(s) = \frac{I_{\text{tank}}(s)}{I_{\text{bias}}(s)} \quad \frac{\frac{g_{m6}}{C}}{s + \frac{g_{m5}}{C}} \tag{9.95}$$

This equation has a dominant pole at

$$P_1 \quad \frac{g_{m5}}{C} \tag{9.96}$$

The behavior of the oscillator must also be determined insofar as it affects the behavior of the loop. We have already shown that the VCO amplitude can be approximated by (9.65). This formula has obvious limitations in describing certain aspects of oscillator performance. Specifically, for large amplitudes, the oscillation amplitude will cease to grow with increasing current and for low current the VCO will not start. More importantly, this expression also fails to capture the frequency response of the oscillator amplitude.

For the purposes of this analysis, the oscillator resonator is treated as a resonator with a pulse of current applied to it by transistors Q_1 and Q_2 each half cycle. From this simple model, the transient behavior of the circuit can be determined. As was shown in (9.1), the resonator forms a time constant $R_p C_{\text{var}}$ that is equivalent to a pole in the response of the oscillation amplitude versus bias current. This pole can be used to give frequency dependence to (9.65).

$$A_2(s) = \frac{V_{\text{tank}}(s)}{I_{\text{tank}}(s)} = \frac{2}{\pi C_{\text{var}}} \quad \frac{1}{s + \frac{1}{R_p C_{\text{var}}}} \tag{9.97}$$

This pole can also be written in terms of the Q and frequency of oscillation:

$$P_2 = \frac{1}{C_{\text{var}} R_p} = \frac{\omega_{\text{osc}}}{2Q} \tag{9.98}$$

It is interesting to note that a resonator with higher Q will respond more slowly and therefore have a lower frequency pole than a low-Q oscillator. This makes intuitive sense, since it is up to the losses in the resonator to cause a change in amplitude.

The last part of the loop consists of the limiting transistors. This is the hardest part of the loop to characterize because by their very nature, the limiters are very nonlinear. The transistor base is essentially grounded, while the emitter is attached to the resonator of the oscillator. In the scheme that has been shown, the base is connected to a voltage higher than the resonator voltage. For any reasonable applied resonator voltage, the current will form narrow pulses with large peak amplitude. Thus, this current will have strong harmonic content. This harmonic content will lead to a nonzero dc current, which is the property of interest. Finding it requires solving the following integral.

$$I_{C\_AVE} = \frac{1}{2\pi}\int_0^{2\pi} I_S e^{\frac{V_{tank}}{2\nu_T}\sin(\theta)}\,d\theta = I_S I_O\left(\frac{V_{tank}}{2\nu_T}\right) \tag{9.99}$$

where $I_O(x)$ is a modified Bessel function of the first kind of order zero, and I_S is the saturation current. For fairly large $V_{tank}/2\nu_T$ there is an approximate solution:

$$I_{C\_AVE} \approx \frac{I_S \times e^{\frac{V_{tank}}{2\nu_T}}}{\sqrt{2\pi\frac{V_{tank}}{2\nu_T}}} \tag{9.100}$$

Now we can write the gain of this part of the loop, which is a nonlinear function of V_{tank} (all other parts of the loop were described by linear functions):

$$A_3(s) = 2\frac{I_{C\_AVE}}{V_{tank}} = \frac{I_S \times e^{\frac{V_{tank}}{2\nu_T}}}{\sqrt{\pi\nu_T V_{tank}}} \approx \frac{I_S \times e^{\frac{V_{tank}}{2\nu_T}}}{\sqrt{\frac{\pi}{\nu_T}}V_{tank}^{\frac{3}{2}}} \tag{9.101}$$

Here it is assumed that this part of the loop has poles at significantly higher frequency than the one in the VCO and the one in the mirror. Note that the value of A_3 will vary depending on the amplitude and thus the loop dynamics may change as the amplitude changes.

These equations can be used to design the loop and demonstrate the stability of this circuit. The capacitor C is placed in the circuit to create a dominant and controllable pole P_1 significantly below the other pole P_2. Generally, the frequency of P_2 and gain of $A_2(s)$ are set by the oscillator requirements and are not adjustable. For more stability, the loop gain can be adjusted by either changing the gain in $A_1(s)$ (by adjusting the ratio of the current mirror) or by adjusting the gain of $A_3(s)$ (this can be done by changing the size of the limiting transistors Q_3 and Q_4). Reducing the gain of the loop is a less desirable alternative than adjusting P_1, because as the loop gain is reduced, its ability to settle to an exact final value is reduced.

Example 9.13: The Design of a VCO AAC Loop

Design an AAC loop for the VCO schematic shown in Figure 9.60. Use $L = 5$ nH ($Q = 10$), $V_{CC} = 5$V. The AAC loop should be set up so that the dc gain around the loop is 40 dB to ensure that the final dc current through the oscillator is set accurately. The loop is to have 0-dB gain at 1 GHz to ensure that parasitic phase shift has minimal impact on the stability of the design. Also, find the phase margin of the loop. Assume that $J_s = 1 \quad 10^{-18}$ A/μm and $\beta_o = 100$ in this technology. Use no more than a gain of 1:10 in the current mirror.

Solution:
We can first compute R_p in the usual way, assuming that the inductor is the sole loss in the system. It is easy to see that at 2 GHz this will be 628.3 . Next, we also need to know the value of the two capacitors in the resonator. Noting that there are two capacitors,

$$C_{\text{var}} = \frac{2}{\omega_{\text{osc}}{}^2 L} = \frac{1}{(2\pi \times 2\text{ GHz})^2 5\text{ nH}} = 2.53\text{ pF}$$

Now since the transistor limiters will turn on for a voltage of about 0.85V in a modern bipolar technology, we can assume that the VCO amplitude will end up being close to this value. Therefore, we can compute the final value for the resonator current from (9.65).

$$I_{\text{tank}} = \frac{V_{\text{tank}}}{0.635 R_p} = \frac{1.7\text{V}}{(0.635)628.3} = 4.26\text{ mA}$$

We will choose to use the 10:1 ratio for the mirror in order to get the best dc gain of 10 from this stage. Thus, the current through the mirror transistor will be 426 μA. We can now compute the values of transconductance and resistance in the model:

$$g_{m5} = \frac{I_{\text{bias}}}{v_T} = \frac{426\ \mu\text{A}}{25\text{ mV}} = 17.04\frac{\text{mA}}{\text{V}}$$

$$g_{m6} = \frac{I_{\text{bias}}}{v_T} = \frac{426\text{ mA}}{25\text{ mV}} = 170.4\frac{\text{mA}}{\text{V}}$$

Once we have chosen the gain for the current mirror, we also know the gain from the resonator:

$$A_2(s) = \frac{V_{\text{tank}}(0)}{I_{\text{tank}}(0)} = \frac{\dfrac{2}{\pi}\dfrac{1}{C_{\text{var}}}}{s + \dfrac{1}{R_p C_{\text{var}}}} = \frac{2}{\pi} R_p = 399\frac{\text{V}}{\text{A}}$$

and since the gain from the mirror will be 10, we can now find the gain required from the limiters.

$$\text{Loop Gain} = 40\ \text{dB} = 20\log(10\times 399\times A_3)$$

$$A_3 = 25.1\frac{\text{mA}}{\text{V}}$$

Thus, from (9.101) we can size the limiting transistors to give the requred gain:

$$A_3 = \frac{I_S e^{\frac{V_{\text{tank}}}{2\nu_T}}}{\sqrt{\pi\nu_T V_{\text{tank}}}} \quad \frac{I_S e^{\frac{V_{\text{tank}}}{2\nu_T}}}{\sqrt{\frac{\pi}{\nu_T}}V_{\text{tank}}^{\frac{3}{2}}} \qquad I_S = A_3 \frac{e^{\frac{V_{\text{tank}}}{2\nu_T}}}{\sqrt{\pi\nu_T V_{\text{tank}}}} \quad \frac{e^{\frac{V_{\text{tank}}}{2\nu_T}}}{\sqrt{\frac{\pi}{\nu_T}}V_{\text{tank}}^{\frac{3}{2}}}$$

$$= 25.1\frac{\text{mA}}{\text{V}} \frac{e^{\frac{1.7\text{V}}{2(25\ \text{mV})}}}{\sqrt{\pi(25\ \text{mV}(1.7\text{V})}} \quad \frac{e^{\frac{1.7\text{V}}{2(25\ \text{mV})}}}{\sqrt{\frac{\pi}{25\ \text{mV}}}1.7\text{V}^{\frac{3}{2}}} = 1.59\times 10^{-17} A$$

This sets the size of these transistors to be 15.9 μm, which is a reasonable size and will likely not load the resonator very much.

Now we can find the pole in the oscillator. It is located at

$$f_2 = \frac{\omega_2}{2\pi} = \frac{1}{2\pi R_p C_{\text{var}}} = \frac{1}{2\pi(628.3\ \ )(2.53\ \text{pF})} = 100.1\ \text{MHz}$$

This means that at 1 GHz, this pole will provide 20 dB of attenuation from the dc value. Thus, the other pole will likewise need to provide 20 dB of attenuation and must also be located at 100 MHz. Since this will lead to both poles being at the same frequency, this design will end up with poor phase margin. We will finish the example as is, and fix the design in the next example.

The other pole is located at

$$f_1 = \frac{\omega_1}{2\pi} = \frac{g_{m5}}{2\pi C} \qquad C = \frac{g_{m5}}{2\pi f_1} = \frac{(17.04\ \text{mS})}{2\pi(100\ \text{MHz})} = 27.12\ \text{pF}$$

This is a large, but not an unreasonable MIM cap, but it could also be made out of a poly cap.

In order to get the phase margin, it is probably easiest to plug the formulas into your favorite math program. You can produce a plot like the one shown in Figure 9.62. The graph shows that this design has about 12° of phase margin (measured as the difference between the phase and 180° at the point where the gain is unity). This is not great. If there is only a little excess phase shift in the loop from something that we have not considered, then this design could very well be unstable. This could be a bad thing. In order to get more phase margin, the poles could be separated or the loop gain could be reduced. Note that the reduction of the loop gain will affect its ability to accurately set the conditions of the oscillator.

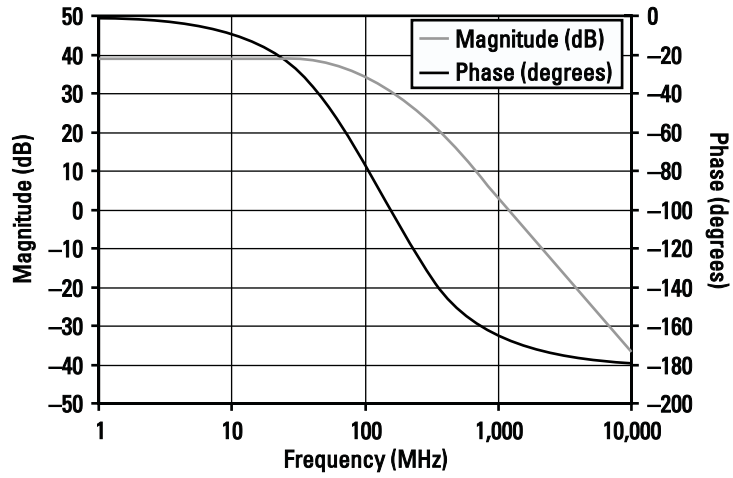

Figure 9.62 Gain and phase response of the AAC loop.

Example 9.14: Improvements to an AAC Loop

Make some suggestions about how to improve stability and simulate a preliminary design.

Solution:
Let us assume that the inductance is still fixed; thus, the pole in the oscillator cannot be adjusted. One obvious thing we can do to make things better is to reduce the loop gain; however, let us start by separating the poles. If we can make the limiting transistors three times smaller at 5.3 μm, but increase the current mirror ratio from 10:1 to 30:1, the gain will remain the same. This will reduce the current in Q_5, which will mean that g_{m5} goes down by a factor of 3. Since the pole is g_{m5}/C, this will move the pole frequency to 33 MHz. We could separate the poles further by increasing C. Note that instead of reducing the pole frequency, this pole could

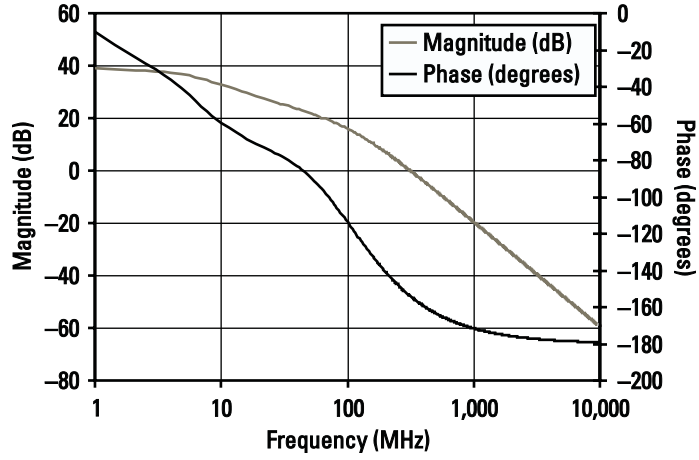

Figure 9.63 Gain and phase response of the improved AAC loop.

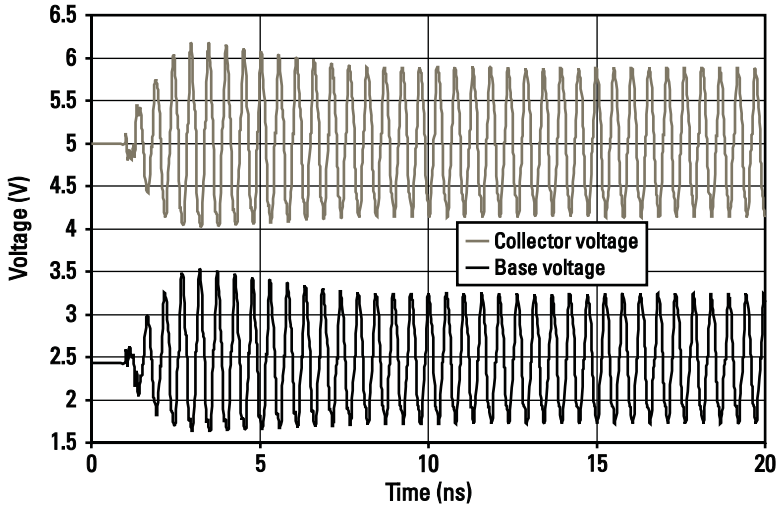

Figure 9.64 Base and collector voltages of the VCO with AAC.

be moved to a much higher frequency (say, 3 GHz), but then the second harmonic would not be attenuated by the loop, which could lead to other problems.

Note that the gain and amount of pole separation are still not enough to make the design practical. If we were designing a product, we would continue to refine these values until we are sure that the feedback loop will not oscillate under any conditions.

We again employ a math tool to plot the frequency response of the equations.

From the plot in Figure 9.63 we can see that there is now about 16° of phase margin, which is still not great, but safer than the original design. Instead of trying to refine the design, let us simply perform some initial simulations to demonstrate the operation of the circuit.

The supply was chosen to be 5V, the base was biased at 2.5V, and the reference current was chosen as 300 μA. The results are plotted in Figure 9.64, which shows the base and collector waveforms of Q_1 or Q_2. Note that the collector voltage is always higher than the base voltage, ensuring that the transistor never saturates.

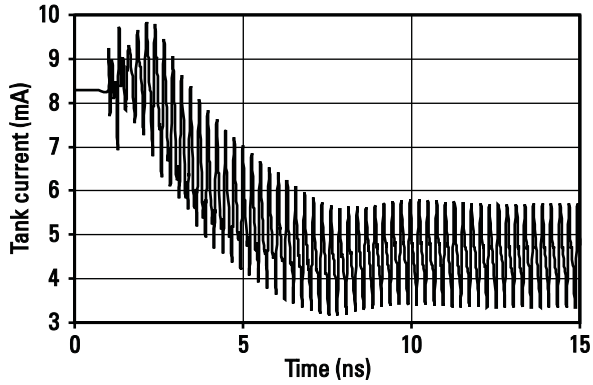

Figure 9.65 Resonator current for the VCO with AAC.

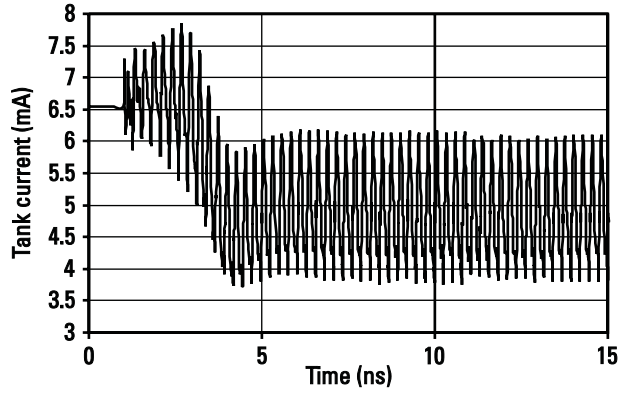

Figure 9.66 Resonator current for the VCO with AAC (reduced phase margin).

Figure 9.65 shows the current in transistor Q_6. Note that we have started the reference current at almost 8 mA. Once the loop begins to operate, it brings this current back to a value necessary to give the designed amplitude. This is about 4.5 mA. Note that this is higher than the estimated current. This is because the resonator voltage is higher than we estimated at the start of the example; thus, more current is required. Figure 9.66 shows the response of the first design biased under similar conditions. Note that with the poles at the same frequency, the response is much less damped. One would expect that in the case for which the poles are widely spaced, the response would look much more first order.

Figure 9.67 shows the differential resonator voltage. Note that it grows and then settles back to a more reasonable value. This peak voltage is slightly higher than that for which it was designed, primarily due to finite loop gain and because we were estimating where the limiting transistors would need to operate. The 1.7-V estimate is quite a rough one. Nevertheless, it served its purpose.

Figure 9.68 is a plot of the current through one of the limiting transistors. This also shows the transient response of the system.

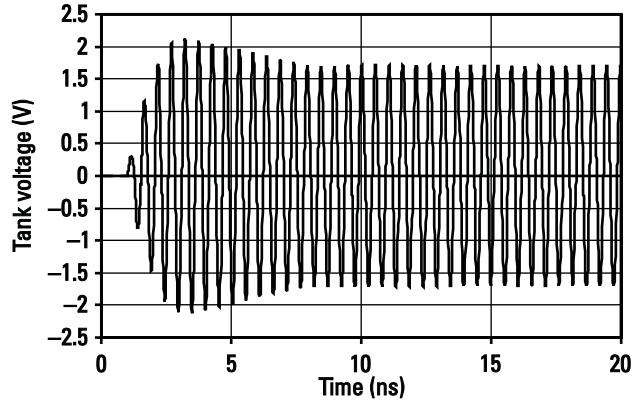

Figure 9.67 Resonator voltage for the VCO with AAC loop.

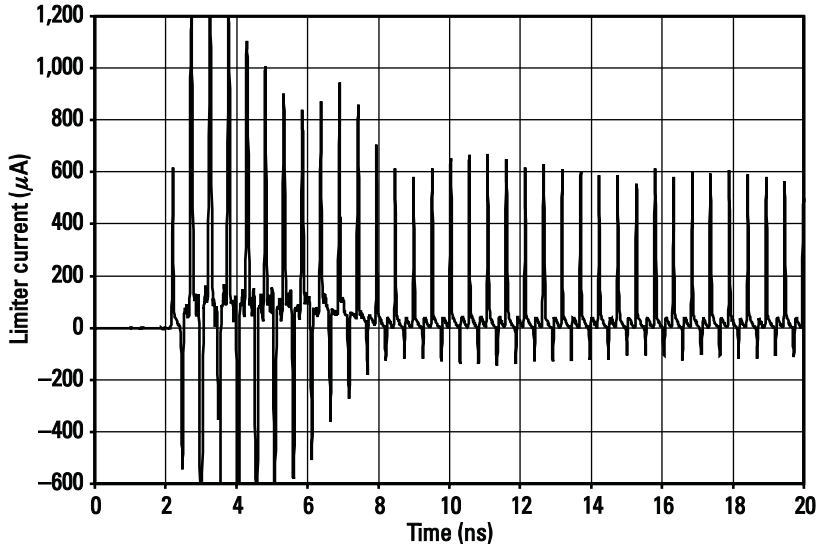

Figure 9.68 Limiting transistor current.

9.21 Supply Noise Filters in Oscillators, Example Circuit

As discussed in previous sections oscillator noise is a very important topic. Often off-chip supply regulators used with IC chips may be insufficient to clean the supply noise to a level where this noise will not dominate the phase noise of the VCO. Remember that in most designs the VCO's varactors will only have a finite isolation from the supply noise. This noise can be very large, and hence can be a dominant source of phase noise. Since VCO performance requires the use of all available headroom, a scheme for filtering the supply noise that does not reduce the supply voltage is needed. A circuit that could be used for this purpose is shown in Figure 9.69.

In this circuit a voltage which is lowpass filtered by the resistor and capacitor is passed through a voltage follower formed by an op-amp and a transistor M_1 to provide the VCO with a supply voltage V_{CC_Clean}. The dc voltage level of V_{CC_Clean}

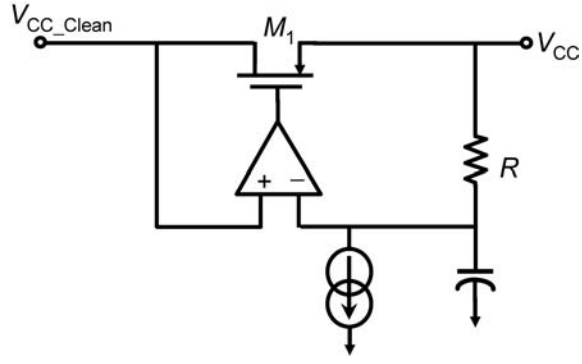

Figure 9.69 A supply filtering circuit.

is set by a current source I_{bias}. It is necessary to move the resistor out of the supply path and use the transistor and op-amp because the resistor's value will have to be quite high to get a filter with a cutoff frequency below 1 MHz, which is what is typically required. This would lead to an unacceptably high voltage drop if the VCO current had to flow through this resistor. The transistor, on the other hand, can be sized to provide the filtering with only a couple of hundred millivolts of drop across it if it is properly designed. Note that the op-amp in this circuit may have to be custom designed to handle input and output voltages very close to the rails.

9.22 Ring Oscillators

Ring oscillators, like all oscillators, must satisfy the Barkhausen criteria for oscillation. However, with ring oscillators it is usually the phase shift that we need to test for, as the gain requirement is usually quite easily satisfied. With a digital style inverter, the gain can be very high at low frequencies. However, the voltage quickly reaches a limit and the effective gain is then zero. Therefore, since the system does not spend its time in the linear region, the concept of gain is a bit more ambiguous than for LC oscillators. However, a common test for startup in ring oscillators is that if you bias all the circuits at their switching points, the loop gain must be bigger than one at the frequency of interest.

A ring oscillator is usually made up of an odd number of inverters or delay cells with the output fed back to the input as shown in Figure 9.70. When power is applied to this circuit, assume the input to 1 is low and the output capacitance C_1 is charged up. When the next stage input sees a high, it will discharge C_2. When the third stage sees a low, it charges C_3. This will in turn discharge stage 1.

Thus, f is related to I/C where I is the charging or discharging current and C is the capacitance size. Thus, frequency can be controlled by changing I. This circuit can operate to high frequency if simple inverters are used. It can be made with CMOS or bipolar transistors. However, ring oscillators, not having inductors, are usually thought to be noisier than LC oscillators, but this depends on the design and the technology.

Now an interesting question at this stage is: How many inverters will be needed to make the circuit oscillate? At first glance, many students may feel that the circuit shown in Figure 9.71(a) should in fact oscillate. If the input is zero, then the output will become one. This is fed back to the input, which then produces a zero at the output, and so on. If the circuit is redrawn as shown in Figure 9.71(b), which is one common way to implement this circuit, it becomes obvious that this is not an oscillator.

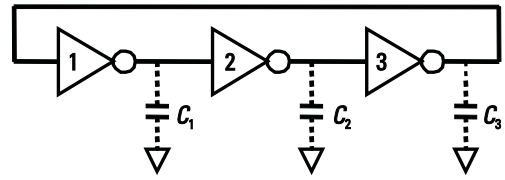

Figure 9.70 A simple ring oscillator.

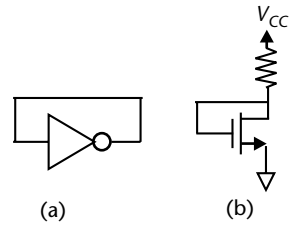

Figure 9.71 An oscillator that is not: (a) block diagram and (b) schematic.

So why, in fact, does it not oscillate? The answer is quite simple; it does not have the required phase shift. The phase shift around that loop is only 180°, which is 180° short of the required 360°.

So the next question is: Can two inverters oscillate? As for any even number of inverters in a ring, this will have 360° of phase shift at low frequency so the circuit will latch and remain in that latched position. In particular, for two stages, the two-inverter ring as shown in Figure 9.72(a) can be redrawn as in Figure 9.72(b), clearly showing that it is equivalent to a latch. For this reason, even numbers of inverters in a ring are avoided.

However, two stages can be made to oscillate if one of the stages is noninverting; for example, with differential circuits, one can simply flip the polarity of the connection between two stages and still have 180° of phase shift for an even number of stages. Then the delays of the circuit make up the additional 180° of phase shift. However, for two stages, assuming each is equivalent to the simple first-order pole shown in Figure 9.73, they each need 90° of phase shift and this will happen only at infinite frequency. Thus, without additional phase shift, oscillation will be very unreliable with two stages. Since ring oscillators with only two stages can operate to higher frequencies (than ring oscillators with more stages), some effort has been put into finding techniques to add additional phase delay for reliable oscillations. Such techniques will be discussed later.

Therefore, in general, the minimum number of stages used to build a ring oscillator is usually three or more. Thus, a three-stage ring oscillator will be the first one that we treat here in more detail. We start by modeling the ring oscillator as three

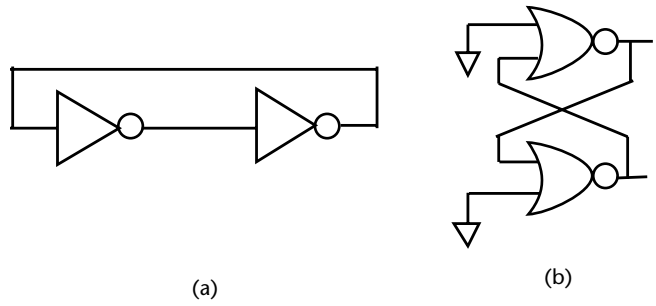

Figure 9.72 Two inverters in a ring: (a) as normally drawn and (b) exactly equivalent circuit of a latch.

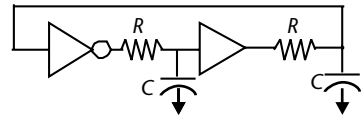

Figure 9.73 Two-stage ring oscillator with delay added at the outputs. Note that this is usually implemented in differential circuits, where one stage has the wires crossed, here represented by a noninverting stage.

negative transconductors driving RC loads as shown in Figure 9.74. Each stage is assumed to have a current gain of G_m from input to output. Therefore, the voltage gain from input to output of one stage is:

$$G(s) = \frac{v_{\text{out}}}{v_{\text{in}}} = \frac{G_m R}{1 + sCR} \tag{9.102}$$

As a result, for three stages the overall open loop gain of the oscillator is:

$$H(s) = \left[\frac{G_m R}{1 + sCR}\right]^3 \tag{9.103}$$

Now each stage contributes a low-frequency phase shift of 180° because of the negative transconductance. At the frequency of interest, there is an additional phase shift per stage of:

$$\phi = \tan^{-1}(RC\omega) \tag{9.104}$$

To obtain a total phase shift that is a multiple of 360°, an additional 180° of phase shift is required from the RC networks so the phase shift in each stage must be equal to 60°. Thus, the frequency of oscillation can be found from (9.104) to be:

$$\omega_{\text{osc}} = \frac{\sqrt{3}}{RC} \tag{9.105}$$

Using this relationship, the loop gain expression (9.103) can be rewritten as:

$$H(j\omega) = \left[\frac{G_m R}{1 + j\sqrt{3}\dfrac{\omega}{\omega_{\text{osc}}}}\right]^3 \tag{9.106}$$

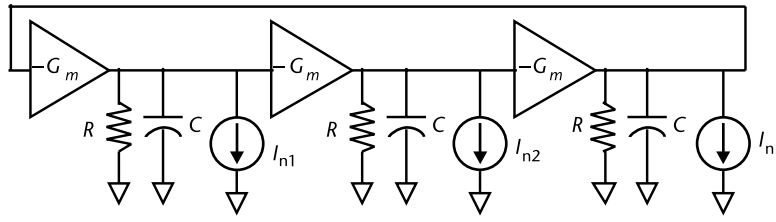

Figure 9.74 Oscillator model also used for noise analysis.

For oscillations in steady state, the loop gain must equal 1. Therefore:

$$\left.\frac{G_m R}{1 + j\sqrt{3}\dfrac{\omega}{\omega_{\text{osc}}}}\right|^{3}_{\omega=\omega_{\text{osc}}} = 1 \tag{9.107}$$

With some manipulation:

$$(G_m R)^3 = 8 \tag{9.108}$$

Therefore:

$$G_m R = 2 \tag{9.109}$$

Similarly, for a four-stage ring oscillator, the required phase shift is 45° per stage. Thus, in this case the frequency is:

$$\omega_{\text{osc}} = \frac{1}{RC} \tag{9.110}$$

and the condition for oscillation is:

$$G_m R = \sqrt{2} \tag{9.111}$$

Using a similar analysis, any order of ring oscillator frequency and gain requirements can be found.

Alternatively, rather than using phase shift, we can characterize an inverter stage by the amount of time delay it has. Thus, for a three-stage oscillator, each stage needs a phase shift of 60° or, equivalently, the delay in each inverter is $T/6$. Consequently, given a delay of τ, the frequency of oscillation will be $1/6\tau$. In general, with an N-stage ring oscillator, the frequency of oscillation is given by:

$$f_{\text{osc}} = \frac{1}{2N\tau} \tag{9.112}$$

Single-ended ring oscillators must be designed with an odd number of stages. However, it is possible to design with an even number of stages if the unit blocks are differential. In this case, one stage is designed to be noninverting by using the opposite outputs from one of the stages. Such an oscillator is shown in Figure 9.75. Now, because there are four stages, it is a simple matter to come up with quadrature outputs as also shown in the diagram.

The output waveform can be assumed to be the result of a constant current I charging up a capacitor C. An estimate of frequency can be made by noting that the output voltage will swing by about half the power supply voltage before triggering the next delay cell. With a power supply voltage V_{DD}, this takes a time of:

$$t = \frac{C}{I}\frac{V_{DD}}{2} \tag{9.113}$$

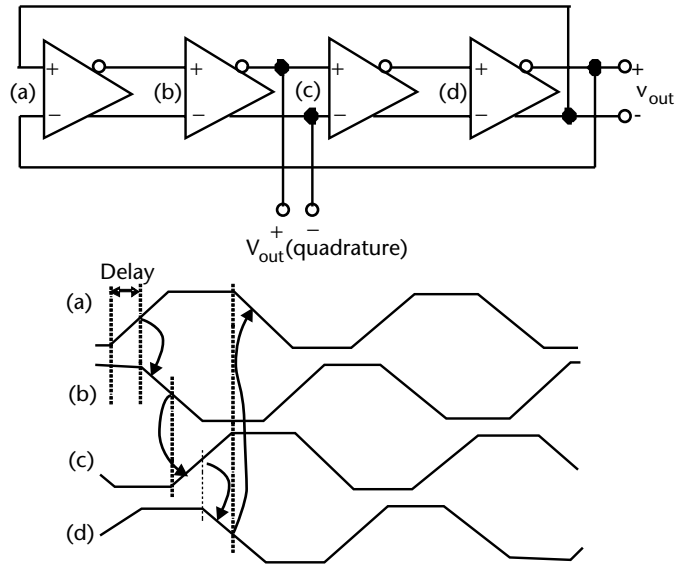

Figure 9.75 A differential ring oscillator with quadrature outputs, with waveforms at each node.

With N stages and noting that (9.113) represents half a cycle, the period can be predicted as follows:

$$T = 2N\frac{C}{I}\frac{V_{DD}}{2} = \frac{NCV_{DD}}{I} \tag{9.114}$$

or equivalently the frequency is given by:

$$f_{\text{osc}} = \frac{1}{T} = \frac{I}{NCV_{DD}} \tag{9.115}$$

As a quick example, if a capacitance as low as 20 fF can be used, to achieve 2 GHz would require a current of about $I = f_{\text{osc}} \cdot N \cdot C \cdot V_{DD} = 2\text{ GHz} \cdot 4 \cdot 20\text{ fF} \cdot 1.2\text{V}$ 200 μA.

Each of the amplifiers or inverters can be made very simply or made in a more complex way. Some examples of simple inverters are shown in Figure 9.76. However, these have no means of tuning. Simple tunable circuits are shown in Figure 9.77.

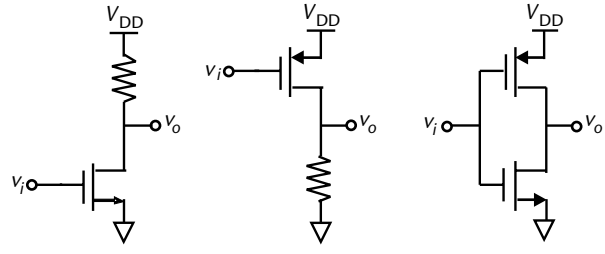

Figure 9.76 Simple single-ended delay cell implementations.

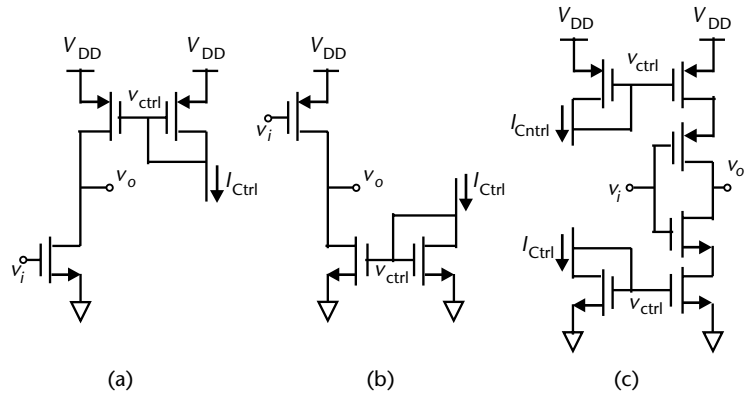

Figure 9.77 Simple tunable single-ended delay cell implementations.

Since oscillating frequency depends on how quickly the interstage capacitance is being charged, the bias current can be used to control the oscillating frequency. To form a differential inverter, it may be tempting simply to combine two of the single-ended inverters, for example, as shown in Figure 9.78. However, this circuit as shown has no connection between the two sides, so there is no reason to believe that the two output signals would be a differential signal. A simple simulation will demonstrate that a four-stage ring oscillator built with this circuit will not work, since the four-stage oscillator depends on the cross coupling of one stage to provide the correct number of inverting elements in the loop. With the cross coupling, since this delay cell does not behave in a differential fashion, this is equivalent to 8 single-ended inverters in a ring, and 8 being an even number causes this circuit to latch up. A second simulation with three inverters in a ring demonstrates that this circuit does work; however, it turns out that generally, the plus and minus rings, being effectively separate rings, will oscillate in phase. Thus, the differential output signal is zero as can be demonstrated by connecting the outputs to a differential amplifier.

Instead, one needs a way to ensure that the two output signals are truly differential, for example, with a differential pair. An example of this is shown in Figure 9.79.

The PMOS transistors and bias voltage can be arranged in several different ways. The first is such that transistors behave like a triode region resistor where the resistance is kept large enough to allow a reasonable output swing. Such a circuit is shown in Figure 9.80(a) in which the PMOS gate is connected to ground. The second way is for the bias voltage and transistor to be designed in such a way that it forms a high impedance current source, an example of which is shown in Figure

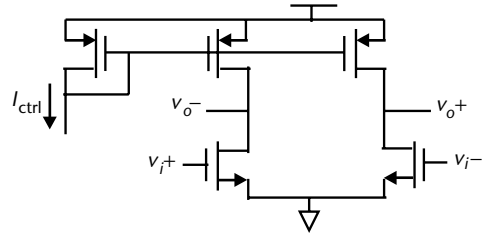

Figure 9.78 A poor differential delay cell implementation.

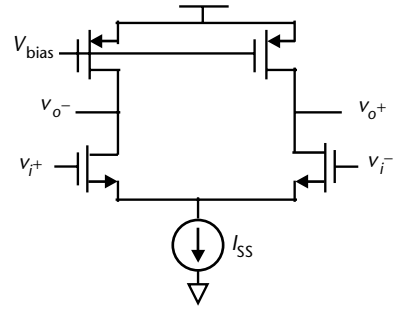

Figure 9.79 Delay cell based on a differential pair.

9.80(b). In such a circuit, the current source I_{SS} must be designed to be larger so it can pull down the output voltage when appropriate. For example, I_{SS} can be made to be twice the current I_3 or I_4 in the PMOS transistors. In Figure 9.80(b), if I_{ctrl1} is equal to I_{ctrl2}, this implies that the NMOS mirror ratio is twice that of the PMOS mirror ratio.

Other variations include using additional diode connected transistors in parallel with the PMOS load transistors, which in [17] is called a symmetrical load. This load arrangement, shown in Figure 9.81, results in a nearly linear transfer function of current versus control voltage. This circuit arrangement was also used by [18] in comparisons with other circuits.

Two circuits with cross-coupled positive feedback are shown in Figure 9.82. The circuit shown in Figure 9.82(a) [19, 20] is a saturated-gain stage with regenerative cross-coupled PMOS transistors which can be controlled to tune the delay (and hence the frequency). This circuit provides for rail-to-rail output signals and full switching of the FETs in the stage. The feedback properties of the latching transistors M_1 and M_2 speed up the signal transitions at the output. The stage also avoids the use of cascode connections and a tail current-source transistor that would limit the signal swing and add more noise to the output. Larger signal swing and faster transitions help to improve signal-to-noise ratio. The circuit in Figure 9.82(b) represents another style of delay cells with cross-coupled transistors providing positive feedback. Variations of this circuit include combining the two bias circuits, adding coarse and fine control of bias currents [21, 22], realization with bipolar delay cells, and adding active inductive loads [23].

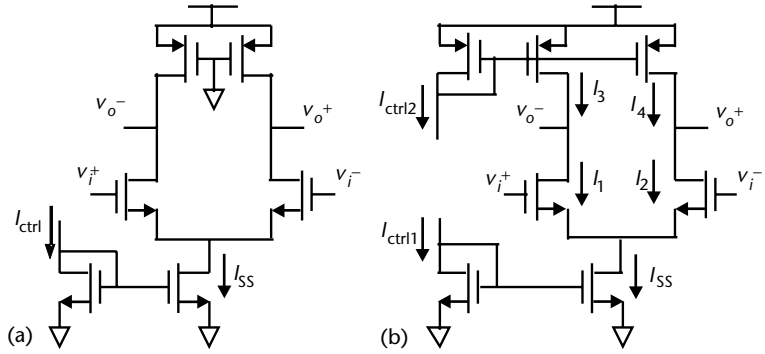

Figure 9.80 (a, b) Delay cell based on a differential pair including biasing details.

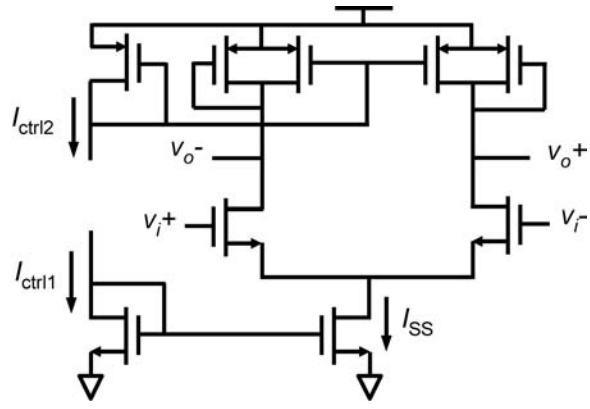

Figure 9.81 Delay cell based on a differential pair with a symmetrical load.

Recently, there have been many designs in which the PMOS current sources are controlled not from the previous stage, but by an earlier stage (being careful to get the polarity right) in order to compensate for the slower speed of PMOS compared to NMOS. While such circuits can be single-ended [24], a differential version is shown in Figure 9.83 [20], in which the delay cell of Figure 9.82(a) has been used.

We note that an oscillator using the circuit in Figure 9.82(a) was compared to an LC oscillator in [25]. The results demonstrated that the LC design was superior in nearly every way; power dissipation was less than 1/10 (3.6 mW compared to 50 mW), phase noise was better by 6 dB when scaled to the same frequency, and the tuning range surprisingly was much better at 46% versus 13.3%. The main advantage for this ring oscillator is that it may have lower layout area since it does not require an inductor. Note that typically ring oscillators would have a wider tuning range than LC oscillators, since LC oscillators are inherently limited by the tuning range of varactors, and ring oscillators have no equivalent fundamental limitation.

Phase noise in ring oscillators is a topic that has been addressed in a number of different ways. In the following it will be analyzed using the same technique used for LC oscillators, that is, to find the effective loop gain and then to calculate the effect of noise introduced into such a loop [7]. Assumption of steady state gain

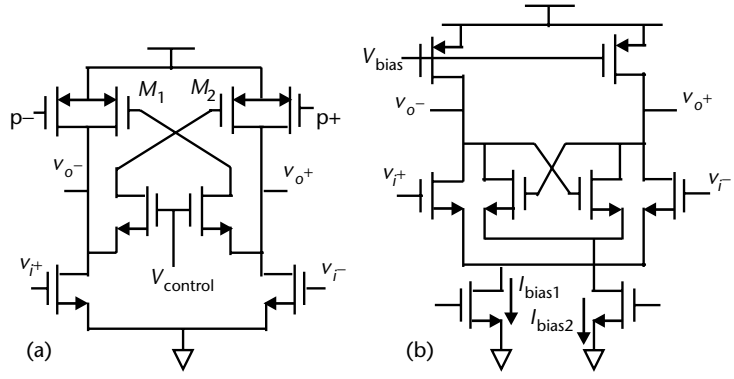

Figure 9.82 (a, b) Delay cell based on differential pair with cross-coupled (positive feedback) to adjust delay.

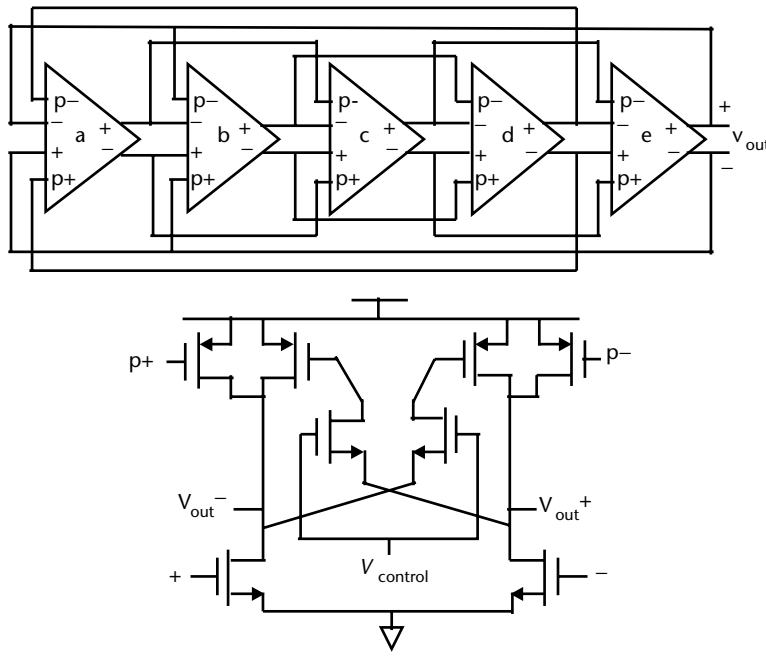

Figure 9.83 Five-stage multiple-pass ring oscillator and saturated gain stage with cross-coupled PMOS transistors.

parameters work quite well for ring oscillators with few stages, as the output tends to be quasi-sinusoidal. For rings with a large number of delay cells, if there is more complete switching, such analyses may be less accurate. More detailed discussion of phase noise and alternative techniques can be found in [10, 26–29]. Starting with the loop gain of a three-stage ring oscillator (as in Figure 9.74) in steady state in (9.105), the frequency of oscillation was shown to be:

$$\omega_{\text{osc}} = \frac{\sqrt{3}}{RC} \tag{9.116}$$

and in (9.106) the loop gain was determined to be:

$$H(j\omega) = \frac{G_m R}{1 + j\sqrt{3}\dfrac{\omega}{\omega_{\text{osc}}}}^{3} \tag{9.117}$$

A further condition in (9.109) was that $G_m R$ was equal to 2. Now if we differentiate $H(j\omega)$ in (9.117) with respect to ω, we get:

$$\frac{dH}{d\omega} = 3(G_m R)^3 \frac{1}{1 + j\sqrt{3}\dfrac{\omega}{\omega_{\text{osc}}}}^{4} j\sqrt{3}\frac{1}{\omega_{\text{osc}}} \tag{9.118}$$

Now we take the magnitude of this expression at ω_{osc}:

$$\left|\frac{dH}{d\omega}\right|^2\Bigg|_{\omega=\omega_{\text{osc}}} = \frac{27(G_m R)^6}{256\omega_{\text{osc}}{}^2} = \frac{27}{4\omega_{\text{osc}}{}^2} \tag{9.119}$$

By combining (9.74) and (9.119), we get:

$$\left|\frac{N_{\text{out}}}{N_{\text{in}}}\right|^2 = \frac{1}{(\ \omega)^2\left|\frac{dH}{d\omega}\right|^2} = \frac{4\omega_{\text{osc}}{}^2}{27(\ \omega)^2} \tag{9.120}$$

Now this is input noise in terms of voltage. Since we have a current:

$$N_{\text{in}} = i_n\ \ \frac{R}{1+j\omega RC}\div \tag{9.121}$$

At the frequency of oscillation:

$$|N_{\text{in}}|^2 = \frac{i_n{}^2 R^2}{4} \tag{9.122}$$

Therefore:

$$\left|\frac{N_{\text{out}}}{i_n}\right|^2_{1\ \text{stage}} = \frac{R^2}{27}\ \ \frac{\omega_{\text{osc}}}{\omega}\div^2 \tag{9.123}$$

With three stages each contributing equal noise:

$$\left|\frac{N_{\text{out}}}{i_n}\right|^2 = \frac{R^2}{9}\ \ \frac{\omega_{\text{osc}}}{\omega}\div^2 \tag{9.124}$$

Now if we approximate the noise coming out of the active circuitry as the drain noise from two CMOS transistors, each with output drain current noise given by:

$$(i_{Dn})^2 = 4\gamma kTg_m \tag{9.125}$$

for long channel devices, the factor γ is approximately 2/3 but for short channel devices, it can be higher, closer to unity. If the two transistors have approximately equal g_m and by noting that $G_m R = 2$, and assuming that small-signal transconductance and effective transconductance are the same ($g_m = G_m$), we come up with the following approximation:

$$i_n{}^2\quad 8kT/R \tag{9.126}$$

Thus, the total output noise is:

$$|N_{\text{out}}|^2 = 8kT\frac{R}{9}\ \ \frac{\omega_{\text{osc}}}{\omega}\div^2 \tag{9.127}$$

Therefore, the phase noise will be equal to:

$$PN(\omega) = \frac{8kT}{v_{osc}^2}\frac{R}{9}\left(\frac{\omega_{osc}}{\omega}\right)^2 \tag{9.128}$$

where v_{osc}^2 is the amplitude of oscillation squared. This is an approximate expression for a three-stage ring oscillator. For a four-stage ring oscillator:

$$\left|\frac{dH}{d\omega}\right|^2\Bigg|_{\omega=\omega_{osc}} = \frac{8}{\omega_{osc}^2} \tag{9.129}$$

and

$$\left|\frac{N_{out}}{i_n}\right|^2 = \frac{R^2}{4}\left(\frac{\omega_{osc}}{\omega}\right)^2 \tag{9.130}$$

for all four stages. Then with the same assumptions about two CMOS transistors, the noise current per stage given by (9.126), the phase noise is given by:

$$PN(\omega) = \frac{2kTR}{v_{osc}^2}\left(\frac{\omega_{osc}}{\omega}\right)^2 \tag{9.131}$$

Comparing (9.131) to (9.128), the phase noise for four stages compared to three stages is 2.25 times or 3.5 dB higher if resistance R is the same. It should be noted that with more stages, the small-signal approximation is less valid and the waveform has a transition portion which could be seen as small-signal or linear; however, each output also spends some time at a nearly constant voltage close to the power supply rails. During the linear portion, the gain is higher than that predicted by $G_mR = 2$, while during the limited portion, the gain is approximately zero. Thus, the linear model, which uses a constant effective gain, is not totally valid. Similarly, the input noise is also cyclostationary. Thus, to get an accurate picture of phase noise, a more complex model needs to be used; however, the above guidelines will give us an idea of the major terms affecting the phase noise. To use this equation to predict phase noise, we find R by using (9.116) where capacitance and frequency of operation are known. For the three-stage ring oscillator, (9.128) can be rewritten as follows:

$$PN(\omega) = \frac{8kT}{v_{osc}^2}\frac{\sqrt{3}}{9\omega_{osc}C}\left(\frac{\omega_{osc}}{\omega}\right)^2 \tag{9.132}$$

Thus, the only variable is the capacitance and given the capacitance, the phase noise can be predicted. So, why are transistor size and bias current not in this equation? We note that they are, indirectly, as they need to be adjusted to give the desired operating frequency. However, once that is done and if the linear approximations are valid, only the capacitance should be important. A similar equation for the

four-stage ring oscillator can be obtained by combining (9.131) with the knowledge that $\omega_{osc} = 1/RC$ resulting in:

$$PN(\omega) = \frac{2kT}{v_{osc}^2 \omega_{osc} C} \left(\frac{\omega_{osc}}{\omega}\right)^2 \quad (9.133)$$

Thus, comparing (9.132) with (9.133) we see that with resistance R removed from the equations, phase noise is about 1.3 times, or about 1.1 dB higher, for a four-stage ring oscillator compared to a three-stage ring oscillator.

In summary, it would appear that any desired phase noise can be achieved simply by choosing an appropriate capacitor size. However, for a larger capacitor, (9.115) shows that more bias current is required to achieve the desired frequency. Thus, there is a direct tradeoff between bias current (and power dissipation) and phase noise. In comparison with LC oscillators, as mentioned in the discussion of Figure 9.82, at the same power dissipation, phase noise is typically worse by 10 to 20 dB. Similarly, to achieve similar phase noise as an LC oscillator, reported ring oscillators have required up to 10 times more power dissipation [25].

Example 9.15: Ring Oscillator Design

Design a single-ended ring oscillator operating at 1 GHz with a phase noise of 100 dBc/Hz at a 1-MHz offset.

Solution:

In a 0.13-μm process, the power supply is typically 1.2V. We will assume the output voltage is 1V peak to peak, or 0.35 V_{rms}. Then for three stages, the phase noise can be predicted using (9.132):

$$PN(\omega) = \frac{8 \quad 4 \quad 10^{-21} \quad \sqrt{3} \quad 2 \times \pi \times 1 \quad 10^{9}}{0.35^2 \quad 9 \times C \times (2 \times \pi \times 1 \quad 10^{6})^2} = 1.27 \quad 10^{-33} \frac{\omega_{osc}}{C} = \frac{8.00 \quad 10^{-24}}{C}$$

A phase noise of 100 dBc is $1 \quad 10^{-10}$ watts. Solving for capacitance, we find:

$$C = \frac{8.00 \quad 10^{-24}}{1 \quad 10^{-10}} = 8.00 \quad 10^{-14} = 80 \text{ fF}$$

Thus, for a little bit of safety, we select a 100-fF capacitor, which would result in a predicted phase noise of 101 dBc/Hz at a 1-MHz offset. This result will not be achieved unless special attention is paid to other details like the minimization of 1/f noise. Otherwise, errors can be 10 dB or more. This will be illustrated in this example. Using 100 fF, we can estimate the current it would take to achieve 1 GHz using (9.115):

$$I = fNV_{DD}C = 1 \quad 10^{9} \quad 3 \quad 1.2 \quad 100 \quad 10^{-15} = 360 \ \mu\text{A}$$

Transistor sizing was done by considering current for maximum f_T. In this process, it is not practical to operate at maximum f_T, since this requires a gate to source voltage close to the power supply voltage. Instead, if the transistor size is increased by about five times, current density is down to about a fifth of the current density for optimal f_T. More importantly, this results in V_{GS} being reduced to about half of V_{DD}, but f_T is only reduced by about 20% of its maximum value so this seems like a reasonable starting point. During the simulation iterations, using the circuit of Figure 9.77(a), the above current was adjusted to 330 μA with a total transistor width of 3 μm with a minimum channel length. Simulation results showed a phase noise of 88.7 dBc/Hz at a 1-MHz offset. Further exploration of phase noise versus capacitance as shown in Figure 9.84 demonstrated that as expected by (9.132) phase noise is inversely proportional to capacitance and proportional to frequency of oscillation.

However, phase noise was considerably higher than predicted by the simple theory by about a factor of 11 dB. A printout of dominant noise sources showed that 1/*f* noise was the main cause of the added phase noise. In an attempt to improve this, transistor sizes (both width and length) were doubled, resulting in phase noise coming down to 94.7 dBc/Hz at a 1-MHz offset, an improvement of about 6 dB. A further doubling of both *W* and *L* resulted in phase noise of 100.8 dBc/Hz at a 1-MHz offset, a further improvement of about 6 dB. Because of the increased parasitic capacitance, it was also necessary to increase the current, in this final case to 470 μA to keep the frequency at about 1 GHz. A simulation of noise versus offset is shown in Figure 9.85. Of importance is to note that the 1/*f* noise corner is at about 700 kHz. This indicates that at a 1-MHz offset, there is still a little bit of room for improvement left.

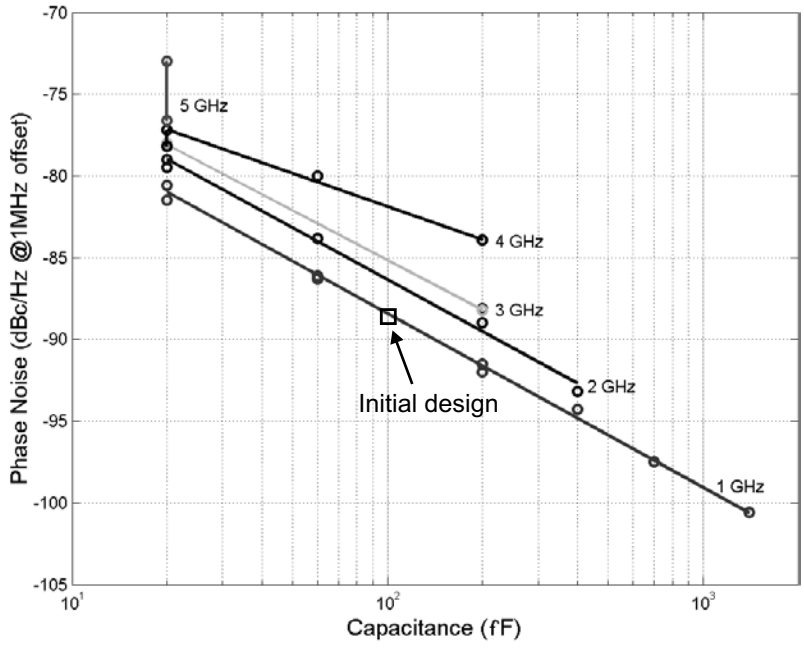

Figure 9.84 Phase noise versus load capacitance for various frequencies. Transistor lengths were minimum, so 1/*f* noise is significant at a 1-MHz offset. Phase noise was further improved by about 12 dB by increasing both *W* and *L*.

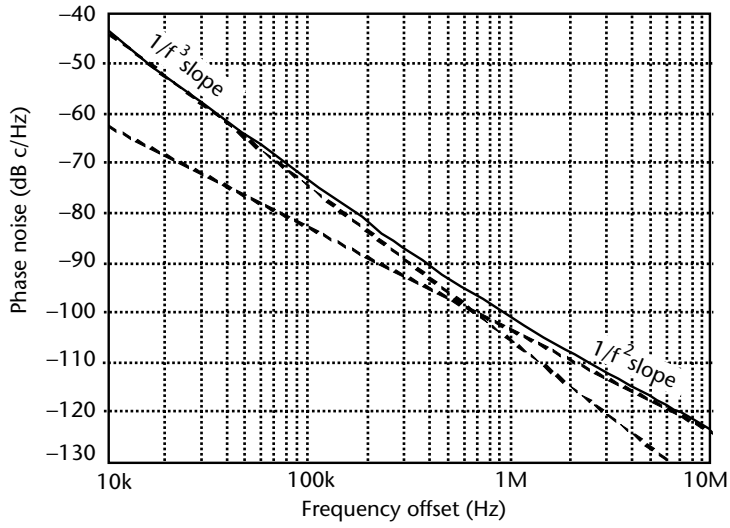

Figure 9.85 Phase noise versus offset frequency showing high 1/*f* corner frequency.

9.23 Quadrature Oscillators and Injection Locking

If, rather than noise, a signal is injected into an oscillator and if that signal is large enough, then it will pull the oscillator to that frequency. This phenomenon is known as *injection locking*. To study the effect of an injected signal, consider the model shown in Figure 9.86(a) where the oscillator feedback is shown on the left and the injected noise and injected signal are shown on the right. The feedback transconductance g_m can be seen as providing a negative resistance R_n equal to $1/G_m$. as shown in Figure 9.86(b). The injected noise current has an expected value of $\sqrt{F \times 4kT/R_p}$ where F, the equivalent noise figure of the oscillator, indicates how much additional noise is added by active circuitry. In addition, it is noted that any input resistance of the transconductance stage has been absorbed in R_p.

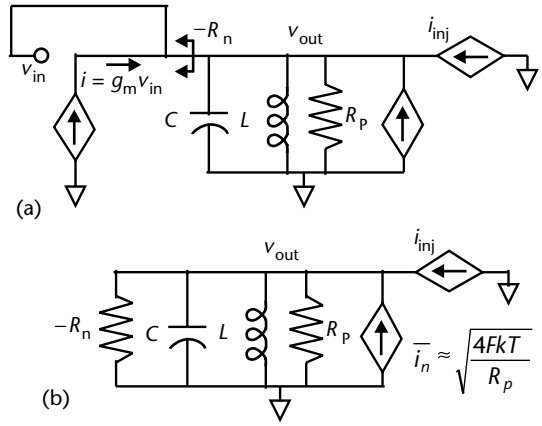

Figure 9.86 Feedback model of oscillator with noise input. This will be used to demonstrate injection locking: (a) with feedback through g_m and (b) with R_n.

Under large signal conditions, the negative and positive resistances in parallel nearly cancel out, resulting in a nearly ideal resonant circuit such that the noise input is amplified to produce the large signal oscillator output voltage v_{out}. Since there is a finite input power, the gain cannot be infinite. However, since the gain is very large, R_n will be approximately equal to R_p. The output voltage is given by:

$$v_{out} = \frac{\overline{i_n}}{\frac{1}{R_p} \quad \frac{1}{R_n} + sC + \frac{1}{sL}} = \frac{s\frac{\overline{i_n}}{C}}{s^2 + s\frac{1}{RC} + \frac{1}{LC}} \tag{9.134}$$

where $1/R = 1/R_p \quad 1/R_n$ or, equivalently, R is the parallel combination of the positive resistor R_p and the negative resistor R_n. If s is replaced by $j\omega$, the following expression results:

$$v_{out} = \frac{\overline{i_n}}{\frac{1}{R} + j \quad \omega C \quad \frac{1}{\omega L} \div} \tag{9.135}$$

The output voltage versus frequency can be seen to be a bandpass filter as shown in Figure 9.87. Resonance occurs at ω_o according to:

$$\omega_o = \frac{1}{\sqrt{LC}} \tag{9.136}$$

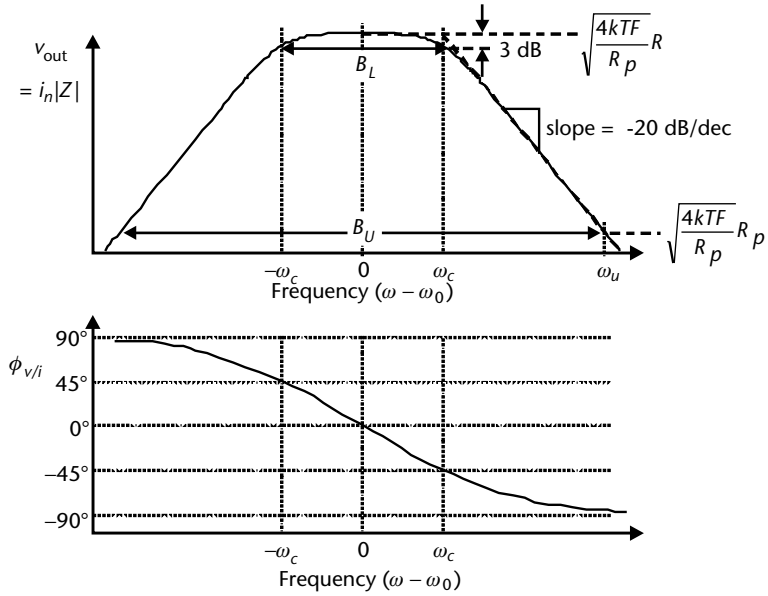

Figure 9.87 The effective gain of oscillator assuming input is thermal noise, and the phase associated with the gain of a bandpass filter.

and the 3-dB bandwidth is given by:

$$B = \frac{1}{RC} \tag{9.137}$$

The output voltage at resonance is given by:

$$v_{\text{out}} = \overline{i_n} R \tag{9.138}$$

When another signal i_{inj} is coupled into an oscillator, whether deliberately or by accident, the output will be the gain times the input signal. If the resulting output signal is larger than the free-running oscillating signal, the oscillator will follow the new signal. Furthermore, in such a case, the oscillator gain will readjust itself (gain will be reduced) so the output amplitude remains approximately constant, and hence the free-running amplitude is the same as the injection locked amplitude. This happens because of the nonlinear limiting mechanism. Note that in general when injection locked, since the gain is reduced, noise will be amplified by a smaller gain; hence, noise is suppressed compared to the free-running case.

To determine the condition to lock, the gain and the free-running amplitude must be determined. From Figure 9.87, the gain A_{osc} can be seen to be simply the net impedance of the equivalent bandpass response as given by:

$$A_{\text{osc}} = \frac{R}{1 + j\dfrac{\omega}{\omega_c}} = \frac{R}{1 + j\dfrac{f}{f_c}} \tag{9.139}$$

The free-running amplitude is determined by the integral under the curve of output voltage and is equal to:

$$\begin{aligned} \sqrt{v_o^2} &= \sqrt{\int |v_{\text{out}}|^2 d\omega} = \sqrt{\int \frac{i_n^2 R^2}{1 + \left(\dfrac{\omega}{\omega_c}\right)^2} d\omega} = \omega_c i_n R \sqrt{\tan^{-1}\left(\frac{\omega}{\omega_c}\right)} \\ &= \overline{i_n} R \sqrt{\frac{B}{2\pi} \times \frac{\pi}{2}} = \frac{\overline{i_n} R \sqrt{B}}{2} = \frac{\overline{i_n}}{2}\sqrt{\frac{R}{C}} \end{aligned} \tag{9.140}$$

It can be seen that $B \cdot \pi/2$ is the noise bandwidth of the bandpass response. Note, that the equation is written for the case where noise current density is given in amperes per $\sqrt{\text{Hz}}$ and B is given in radians/second, hence an extra factor of 2π has been included. If both terms are expressed in the same units, whether hertz or radians per second, the factor of 2π should be removed from (9.140).

The amplitude of oscillation is typically determined by the nonlinear limiting of transconductor g_m. For small signals, the value of g_m should result in a net negative resistance and an unstable circuit. With increasing amplitude, limiting causes the value of g_m to decrease until $|R_n| \quad |R_p|$ for steady-state operation. Transconductance g_m is defined by the current voltage relationship as follows:

$$i = k_1 v + k_2 v^2 + k_3 v^3 + \cdots \tag{9.141}$$

For small signals, k_1 can be seen as g_m. However, for larger signals the third-order term will produce components at the fundamental frequency and if k_3 is negative will result in a decrease in the effective value of g_m. Specifically, if $v = V\cos(\omega t)$:

$$i = k_1 V + \frac{3k_3 V^3}{4}\cos(\omega t) + \frac{1}{2}k_2 V^2 \cos(2\omega t) + \frac{1}{4}k_3 V^3 \cos(3\omega t) + \cdots \tag{9.142}$$

Thus, the effective g_m is given by the fundamental component of $|i/v|$:

$$g_m = k_1 + \frac{3k_3 V^2}{4} \tag{9.143}$$

This can be solved for amplitude V as:

$$V = \sqrt{\frac{4}{3k_3}\left(\frac{1}{R_p} - k_1\right)} \tag{9.144}$$

We note that k_1 is typically larger than $1/R_p$ to insure that the oscillator starts up. However, this means the last term in the brackets of (9.144) is negative, but since k_3 is typically also negative, the square root can be taken. Equations (9.139) and (9.140) can be combined to determine the minimum signal amplitude for which the output due to the injected signal is equal to or larger than the output amplitude due to the noise.

$$\left|\frac{i_{\text{inj}}R}{1 + j\frac{f - f_o}{f_c - f_o}}\right| > \frac{\overline{i_n}R\sqrt{B}}{2} = v_o \tag{9.145}$$

Here frequencies are expressed as offsets with respect to the free-running frequency (see Figure 9.87). Note once again that the original noise input signal is a current density, in amperes per $\sqrt{\text{Hz}}$, while B is in units of radians per second. For an injected signal inside of the oscillator bandwidth, the condition for lock is:

$$i_{\text{inj}} > \frac{\overline{i_n}\sqrt{B}}{2} = \frac{v_o}{R} \tag{9.146}$$

For injected signals sufficiently outside the corner of the oscillator bandwidth, the condition for locking is:

$$i_{\text{inj}} > \frac{\overline{i_n}(f - f_0)\sqrt{B}}{2(f_c - f_0)} = \frac{\overline{i_n}(f - f_0)\sqrt{B}}{2\frac{B}{2}\frac{1}{2\pi}} = \frac{\overline{i_n} \times 2\pi(f - f_0)}{\sqrt{B}} = \frac{v_o}{R}\frac{(f - f_o)}{(f_c - f_o)} \tag{9.147}$$

where $(f \quad f_0)$ is the frequency offset from the oscillator free-running frequency. Thus, for larger offsets, the injected signal needs to be stronger. Note that an oscillator can also injection lock to a harmonic of a signal, and the above analysis can be used to determine the required amplitude of the harmonic signal. For example, a free-running oscillator at 1.1 GHz can be made to lock on to the eleventh harmonic of a 100-MHz input signal, provided that the input signal is nonsinusoidal enough such that it has sufficient amplitude at the eleventh harmonic. In this example, since the eleventh harmonic is close to the free-running frequency, lock will be achieved with a small input signal. Any other harmonic being far away from the free-running frequency would require much larger signal amplitudes to lock.

A loop analysis can be used to find an alternative expression for oscillation amplitude due to injected current i_{inj}. Input current i_{inj} can be expressed as an equivalent voltage v_{inj} applied at the input of the transconductor by dividing current i_{inj} by g_m. Then the output voltage is found from the forward gain FG in terms of the transconductance g_m and the open-loop conductance Y_{OL} as follows:

$$v_{\text{out}} = v_{\text{inj}} \frac{FG}{1 \quad FG} = \frac{i_{\text{inj}}}{g_m} \frac{\dfrac{g_m}{Y_{OL}}}{1 \quad \dfrac{g_m}{Y_{OL}}} = i_{\text{inj}} \frac{1}{Y_{OL} \quad g_m} = \frac{i_{\text{inj}}}{\dfrac{1}{R_p} + j\omega C \quad 1 \quad \left(\dfrac{\omega_o}{\omega} \div \div \right)^2 \quad \dfrac{1}{R_n}} \tag{9.148}$$

Not surprisingly, the result can be shown to be the same as previously shown in (9.134). We can manipulate the expression for Y_{OL} by noting that the unloaded quality factor Q_U is equal to ω_o/B, and bandwidth B is equal to $1/R_pC$. As a result, the following expression for open-loop admittance Y_{OL} can be obtained.

$$Y_{OL} = \frac{1}{R_p} + j\omega C \quad 1 \quad \left(\frac{\omega_o}{\omega} \div \div \right)^2 = \frac{1}{R_p} \quad 1 + jQ_U \frac{\omega}{\omega_o} \quad 1 \quad \left(\frac{\omega_o}{\omega} \div \div \right)^2 \tag{9.149}$$

If ω is replaced by $\omega_o + \quad \omega$, and the approximation is made that $\quad \omega$ is much less than ω_o, the following approximation is obtained:

$$Y_{OL} = \frac{1}{R_p} \quad 1 + jQ_U \frac{\omega}{\omega_o} \quad 1 \quad \left(\frac{\omega_o}{\omega_o + \quad \omega} \div \right)^2 \quad \frac{1}{R_p} \quad 1 + j2Q_U \frac{\omega}{\omega_o} \tag{9.150}$$

Substituting back into the original expression:

$$v_{\text{out}} = i_{\text{inj}} \frac{1}{Y_{OL} \quad g_m} \quad \frac{i_{\text{inj}}}{\dfrac{1}{R_p} \quad 1 + j2Q_U \dfrac{\omega}{\omega_o} \quad g_m} = \frac{i_{\text{inj}} R_p}{1 \quad g_m R_p + j2Q_U \dfrac{\omega}{\omega_o}} \tag{9.151}$$

It is interesting to note that this has been interpreted to mean that the output voltage depends on the original parallel resistance and hence Q of the tank circuit. However, R_p and Q_U can both be eliminated from this expression by noting that since $|R_p|$ is approximately equal to $|R_n|$, $1 \quad g_m R_p$ is very small. Thus, the expression for output voltage (which could have been derived simply enough directly from the original expression) is given by:

$$v_{\text{out}} \quad \frac{i_{\text{inj}}}{j2C \quad \omega} \tag{9.152}$$

This very simple result, which can be used as an alternative to (9.147), seems to show that the original Q_U and parallel resistance R_p are irrelevant. This can be explained by noting that feedback produces negative resistance, which exactly cancels the original positive resistance. However, one should not be too hasty in stating that original R_p and Q_p are irrelevant, since the noise current i_n has a component directly related to R_p or equivalently Q_U, and depending on the original formulation of the equivalent circuit, the factor of g_m can appear in the expression for v_{out}.

9.23.1 Phase Shift of Injection Locked Oscillator

From the model of the oscillator in Figure 9.86, i_{total}, the total current injected into the resonant circuit, is the vector sum of i and i_{inj}. If the oscillating frequency is not at the center frequency of the bandpass filter, there will be phase shift between the total current injected into the tank and the voltage it produces as given by:

$$\phi_{\text{osc}} = \quad \tan^{\ 1} \quad \omega C \quad \frac{1}{\omega L} \div R \tag{9.153}$$

As noted before, the value of R is ultimately set by the nonlinear limiting, typically by the transconductor. With an injected signal, the output voltage is still determined from the integral as in (9.140), however, with the additional term of $i_{\text{inj}}\ R$ added to it. Since the output voltage remains roughly constant, it is clear that for larger injected signals, the value of R must decrease; hence, the gain to the noise also decreases. As well, as seen from (9.153), with a decrease in R, there is less phase shift for a given frequency offset.

The phase in (9.153) is the total phase and is the combination of two components. The first is the phase between i and v. Since i is directly created from the voltage, it must be in phase with the voltage. Hence, there must be additional phase shift between i_{inj} and the voltage such that the total current has the correct phase shift with respect to the voltage as shown in Figure 9.88. In the general case, the phase shift can be calculated from:

$$\phi_{\text{inj}} = \sin^{\ 1} \quad \frac{i_{\text{total}}}{i_{\text{inj}}} \sin \phi_{\text{osc}} \div \tag{9.154}$$

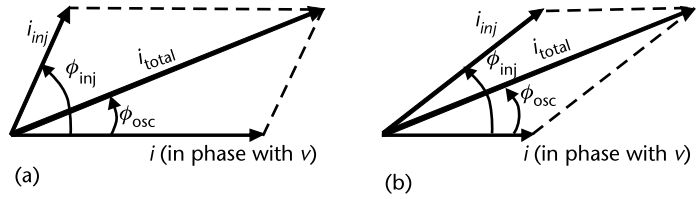

Figure 9.88 Phase shift of an external signal i_{inj} with respect to the output voltage: (a) for a small injected signal and (b) for a larger injected signal.

Note that if the amplitudes of current were the same, (as will be the case in the quadrature oscillator to be discussed in a following section) then the phase shift ϕ_{inj} would be twice that required by the bandpass filter ϕ_{osc}.

$$\phi_{inj} = 2\phi_{osc} = \quad 2\tan^{1} \quad \omega C \quad \frac{1}{\omega L} \div R \tag{9.155}$$

In the special case where phase is 90°, the filter has a phase shift of 45° that occurs at:

$$\omega_{90^\circ} = \sqrt{\omega_0{}^2 + \frac{B^2}{4}} \pm \frac{B}{2} \quad \omega_0 \pm \frac{B}{2} \tag{9.156}$$

where the approximation is valid if the center frequency is much larger than the bandwidth.

Note that in practice, deciding what value of R to use in the above equations to predict phase shift is not obvious, since R changes with the amount of injected current. From experience, and somewhat surprisingly, it has been observed that using the parallel tank resistor R_p for R gives good results.

Two oscillators can be connected in such a way that a signal from one oscillator is injected into the second oscillator and a signal from the second oscillator is injected into the first. The result is that the two oscillators become locked in frequency, typically oscillating in quadrature. Very often quadrature signals are required, for example, mixers in an image reject configuration require quadrature oscillator signals.

9.23.2 Parallel Coupled Quadrature LC Oscillators

The most common technique is the parallel connection as shown in Figure 9.89 where each oscillator is made up of a tank circuit and cross-coupled feedback circuit. In addition, each oscillator output is connected to the other oscillator with transistors in parallel to the cross-coupled transistors. Thus, oscillator 1 has feedback transistors M_1 and M_2 and coupling from oscillator 2 via transistors M_5 and M_6. Typically, feedback and coupling transistors are made the same size. Furthermore, because of symmetry, the oscillation amplitudes of the two oscillators should be the same.

To understand why the two oscillators oscillate in quadrature, first note that the two oscillators can be modeled as gain stages, as shown in Figure 9.90. Each stage

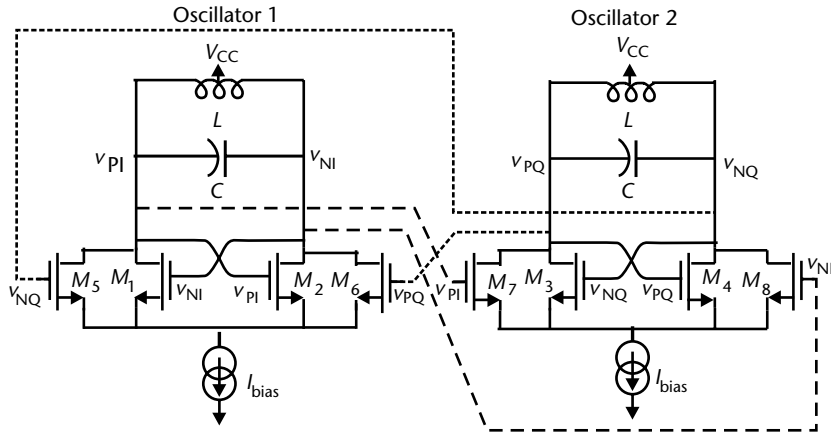

Figure 9.89 Quadrature negative G_m oscillator with parallel cross connections.

has a gain and a phase as shown in Figure 9.87. We note that, as for any oscillator structure, the loop phase must be 0° or 360°. Thus, since the crossed wires at the output represent a phase shift of 180°, the two oscillators together must have an additional phase shift of 180°. Hence, if all components are matched, the phase shift across each oscillator will be 90°. We note that Figure 9.87 shows the transfer function to the total current injected into the tank, and for such current, the phase only asymptotically approaches 90°. The total current is made up of a combination of two equal currents as shown in Figure 9.91.

The first current i is locally fed back in direct proportion to the voltage; hence, it is in phase with the voltage. The second component i_{inj} comes from the other oscillator. Since the two currents have the same amplitude, for the two oscillators to be 90° out of phase, the total current is 45° out of phase with respect to the voltage. By this argument, it can be seen that the frequency of operation will be shifted from the resonant frequency by an amount of $B_L/2$, shown in Figure 9.87, since this is where there is 45° of phase shift.

Phase shift in the injection-locked oscillator was previously given by (9.155) and here, repeated as (9.157), derived with equal currents for feedback and cross coupling.

$$\phi_{\text{inj}} = 2\,\phi_{\text{osc}} = \quad 2\,\tan^{\quad 1} \quad \omega C \quad \frac{1}{\omega L} \div R \tag{9.157}$$

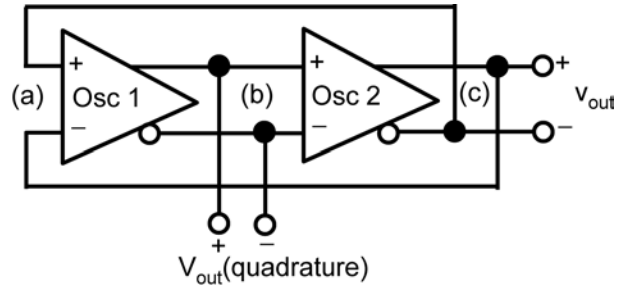

Figure 9.90 A quadrature oscillator modeled as two amplifier stages in feedback.

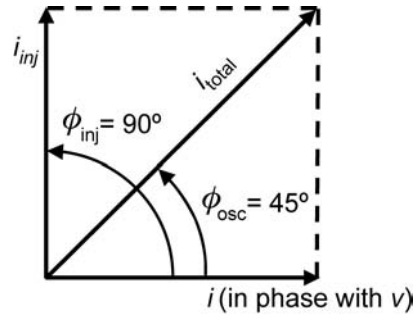

Figure 9.91 The current injected into each resonator for a quadrature oscillator.

In the special case where phase is 90°, the filter has a phase shift of 45° that occurs at:

$$\omega_{90^\circ} = \sqrt{\omega_0{}^2 + \frac{B^2}{4}} \pm \frac{B}{2} \quad \omega_0 \pm \frac{B}{2} \tag{9.158}$$

where the approximation is valid if the center frequency is much larger than the bandwidth. In practice, these equations work well if R_p is used for R.

An analysis of the loop gain of the quadrature oscillator model in Figure 9.92 results in:

$$\frac{v_1}{v_1} = \frac{g_m{}^2}{g_m \quad \frac{1}{R} \div + j \quad \omega C \quad \frac{1}{\omega L} \div}{}^2 \tag{9.159}$$

If it is assumed that each oscillator in steady state has adjusted its g_m such that it is equal to (or close to) the value of $1/R$, and if the loop gain is set equal to 1, then:

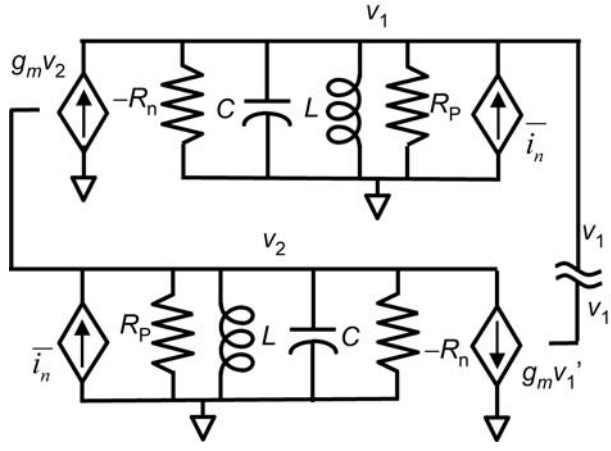

Figure 9.92 A model for a quadrature oscillator.

$$\omega^2 LC \pm \omega L g_m \quad 1 = 0 \tag{9.160}$$

The plus or minus of the middle term is used because one can find two solutions. One solution is below the resonant frequency, and one is above the resonant frequency. The solution for ω is:

$$\omega = \sqrt{\frac{1}{LC} + \frac{gm^2}{4C^2}}^2 \pm \frac{g_m}{2C} \quad \omega_0 \pm \frac{B}{2} \tag{9.161}$$

where B is the equivalent bandwidth and is given by g_m/C.

Thus, although some approximations have been made, the result is that the circuit behaves like two coupled parallel resonant circuits with equivalent bandwidth of g_m/C. If it is noted that g_m is nominally equal to $1/R_p$, then the frequency of oscillation is in exact agreement with the previous determination. However, the phase shift does not agree exactly at frequencies other than the resonant frequency, but if such phase information is desired, (9.157) can be used.

Example 9.16: Quadrature Oscillator Design

Design a 1-GHz quadrature oscillator using simplified models to demonstrate phase shift and amplitude theory. Change the capacitance by 1%, and observe and explain the resulting amplitude and phase mismatch.

Solution:

A 1-GHz quadrature oscillator was built with a resonant tank made up of 5 nH and 5.06 pF. Feedback and cross-coupling transconductors had transfer functions $i = 0.005v \quad 0.0005v^3$. Initially, the circuit was run open loop with a voltage representing VCO_2, with that voltage adjusted to be equal to the oscillating amplitude of VCO_1, and magnitude and phase results were obtained as in Figure 9.93.

Thus, where amplitude is rolled off by 3 dB (down to about 2.1V), phase is at 90°, as expected from the explanation around Figure 9.91. When connected as an injection-locked oscillator, the oscillating frequency was 1.05144 GHz, at the frequency where the phase shift is 90° with amplitude of about 2.2V, also as expected. The phase shift between the two oscillators was exactly 90° within the simulation limits (better than a thousandth of a degree). When capacitance of one oscillator was increased by 1%, the frequency decreased to 1.0487 GHz and the phase shift was now 93.93°. This can be explained by examining a zoom in of the phase versus frequency plot derived from (9.157) and shown in Figure 9.94. When both capacitors are at 5.06 pF, each phase shift is 90° at a predicted frequency of 1.0519 GHz, close to the simulated frequency of 1.05144 GHz. When one capacitor is high by a fraction δ (in this case 1%) at 5.1106 pF, its resonant frequency decreases by about $\delta/2$ (in this case by 0.5%), and consequently, there is more phase shift at the frequency of interest. The total phase still has to be 90°; hence, frequencies adjust themselves until the sum of the two phase shifts is 180°. This new frequency can be found by noting where the average of the two phase shifts goes through 90°

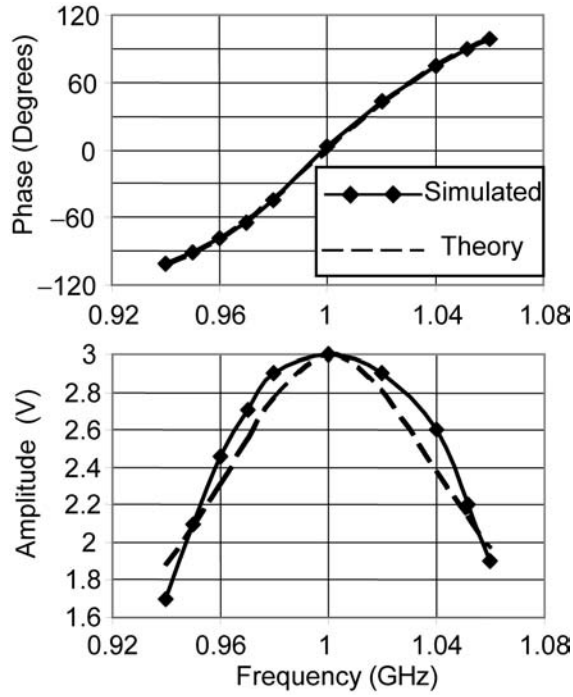

Figure 9.93 Quadrature oscillation phase and amplitude for an injected signal (open-loop simulation).

approximately at a frequency shift of $\delta/4$ (or 0.25%). From the starting frequency of 1.0519 GHz, a 0.25% shift would move it to 1.0493 GHz, while the equation and Figure 9.94 predict a new frequency of 1.0492 GHz. Both are close to the simulated frequency of 1.0487 GHz. Also of importance, the phase shift across the two oscillators is now seen to be about 87° and 93° for a total phase of 180°,

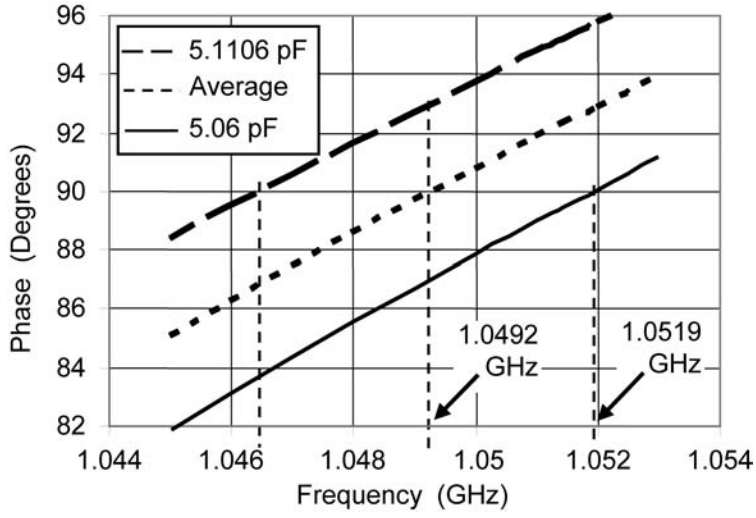

Figure 9.94 Phase of an injected signal for a quadrature oscillator with nominal capacitor of 5.06 pF and 1% increased capacitor of 5.1106 pF (open-loop simulation).

again in agreement with the simulations. Phase shift can be shown to be related to bandwidth by:

$$\left|\frac{d\phi}{d\omega}\right| = \frac{2Q}{\omega_o} = \frac{2}{B} \tag{9.162}$$

Hence, the phase shift is estimated at:

$$\phi = \frac{\omega}{B/2} = \frac{\omega_{\text{matched}} \times \delta/4}{B/2} \tag{9.163}$$

where δ is the capacitor mismatch (0.01), ω_{matched} is the quadrature oscillator frequency with components matched (2π 1.0519 G) and $B/2$ is the difference between the LC resonant frequency and the free-running frequency (2π 51.6 M). This equation predicts a phase offset of 2.92°, which is quite close to the value predicted from Figure 9.94. The simulated phase change is slightly larger at 3.93° but still illustrates the usefulness of this estimate.

Similarly, the change of phase shift is also related to a change of amplitude as was seen in Figure 9.93. Simulated results show the amplitudes are now 2.08V and 2.35V in agreement with the above explanation.

9.23.3 Series Coupled Quadrature Oscillators

Quadrature oscillators can also have series coupling as shown in Figure 9.95. These two topologies have been studied and compared by [30]. A conclusion was that optimal coupling for the parallel circuit resulted when the coupling and main

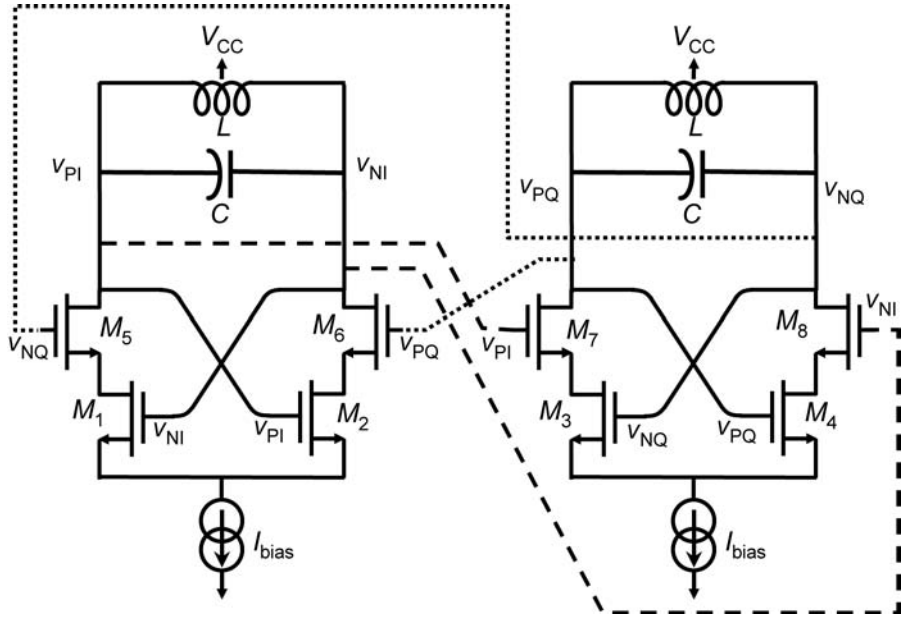

Figure 9.95 A quadrature negative G_m oscillator with series cross connections.

transistor were approximately the same size. For the series circuit, the coupling transistors should be about five times bigger than the main transistors. For optimal coupling, the parallel circuit appeared to have better quadrature phase matching; however, if its coupling transistor size was adjusted so that the quadrature phase matching was equal, then the series coupled circuit was stated to have 10 dB to 20 dB better phase noise.

9.23.4 Other Quadrature Generation Techniques

All of the quadrature schemes shown so far have some drawbacks. For example, the series-coupled oscillator requires a coupling transistor about five times bigger than the other transistors. These coupling transistors will have significant parasitic capacitance, which will limit tuning range. The parallel-coupled circuit has additional power dissipation due to the coupling transistors and has less than ideal phase noise. In both of the above coupling techniques, since the injected signal is being applied at 90° to the oscillating signal, the peak injected current occurs at the zero crossing of the oscillator signal, and at this point, the oscillator is most sensitive to injected noise. It has been shown in [31] and others that by applying 90° of phase shift in the coupling path, the injected signal can be in phase with the oscillating signal, and phase noise is substantially improved. However, this improvement comes at the cost of needing further resonator circuits, and does not solve the other problems noted above (increased power dissipation for parallel coupling and reduced tuning range for series coupling). An alternative way to improve phase noise is to couple the two oscillators by use of the voltage on the current source node as shown in Figure 9.96. Since this voltage is typically at a harmonic of the oscillating frequency, this technique is referred to as superharmonic coupling [32]. This technique is lower power than the parallel technique, as no extra current paths are required.

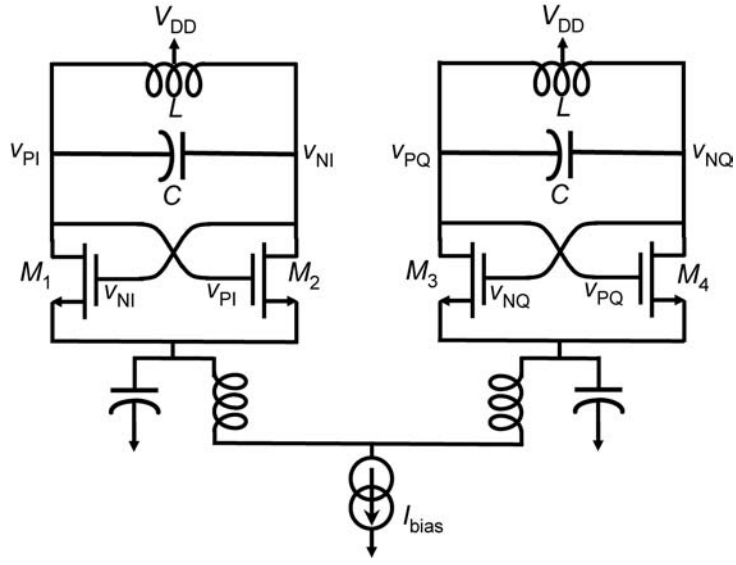

Figure 9.96 A quadrature negative G_m oscillator using superharmonic coupling.

9.24 Other Oscillators

Although we have stressed LC-based VCOs in this chapter as the most common RF oscillators due to their excellent phase noise, there are many other ways to build a circuit that generates harmonic waveforms. For example, we have already discussed the ring oscillator. In this section we will briefly look at the multivibrator, and then examine the crystal oscillator in more detail.

9.24.1 Multivibrators

The voltage-controlled, emitter-coupled multivibrator oscillator is a circuit that alternately charges and discharges a capacitor to produce a periodic waveform. The frequency is set by the capacitor size and the current. Since current can be programmed over a wide range, this oscillator can have a broad frequency tuning range. The maximum oscillating frequency is limited by the minimum reasonable capacitor size and the speed of switching. As in digital circuitry, the highest switching speed would typically be achieved with emitter-coupled logic or current-mode logic. Since only transistor based current sources and capacitors are required, this oscillator is inherently compact. However, typically LC oscillators or ring oscillators are preferred for RF frequencies; hence, this oscillator will not be further discussed.

9.24.2 Crystal Oscillators

Quartz crystal resonators are widely used in frequency control applications because of their unequaled combination of high Q, stability, and small size. When a potential difference is applied across opposite faces of a quartz crystal, mechanical deformation takes place. If the frequency of the potential is appropriate, the crystal will vibrate and indeed resonate. The resonant frequency, Q, and temperature coefficient depend on the physical size and the orientation of faces relative to the crystal axis. The resonators are classified according to "cut," which is the orientation of the quartz wafer with respect to the crystallographic axes of the material. Examples are AT-, BT-, CT-, DT-, and SC-cuts, but they can also be specified by orientation, for example, a +5° X-cut. Although a large number of different cuts have been developed, some are used only at low frequencies, others are used in applications other than frequency control and selection, and still others have been made obsolete by later developments. At frequencies above approximately 1 MHz, AT- and SC-cuts are primarily used.

For most applications, the two-terminal equivalent circuit consisting of the static capacitance C_0 in parallel with the dynamic or motional branch, L_1-C_1-R_1, is used as shown in Figure 9.97, in which f_s is called the motional resonance frequency given by:

$$f_s = \frac{1}{2\pi\sqrt{L_1 C_1}} \tag{9.164}$$

For some applications, harmonics, or overtones, are used, in which case the model has more branches in parallel, one for each harmonic.

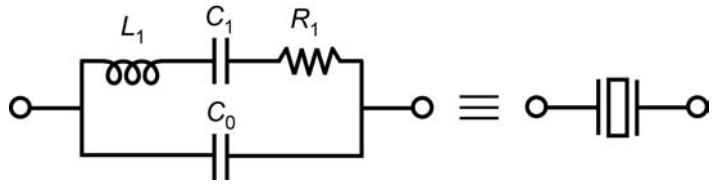

Figure 9.97 Two-terminal equivalent circuit of a crystal.

For oscillator applications, the figure of merit, M, is a useful indicator that is defined as:

$$M = \frac{1}{2\pi f_s C_0 R_1} \tag{9.165}$$

For M less than 2, the crystal reactance is never inductive at any frequency, and an additional inductor would be required to form an oscillator. In general, a larger M results in a more useful resonator.

In a crystal resonator, the quality factor of a reactive component is the reactance X_1 of the motional inductance or capacitance divided by the motional resistance R_1:

$$Q = \frac{|X_1|}{R_1} = \frac{2\pi f_s L_1}{R_1} = \frac{1}{2\pi f_s C_1 R_1} \tag{9.166}$$

where the time constant $\tau = C_1 R_1$ depends on the mode of vibration and on the angles of cut. For AT-cut c-mode, $\tau = 10$ fs, for SC-cut c-mode, $\tau = 9.9$ fs, and for BT-cut b-mode, $\tau = 4.9$ fs [33]. Practically, quartz crystal resonators can have an unloaded Q up to more than a million and a temperature drift of less than 0.001% over the expected temperature range. The maximum Q which can be obtained is determined by several additive loss factors, the first of which is the intrinsic Q of quartz, which is approximately 16×10^6 divided by the frequency in megahertz for the AT-cut and slightly higher for the SC-cut. Other factors that further limit Q are mounting loss, atmospheric loading (for nonevacuated crystal units), and the surface finish of the blank. Mounting loss depends upon the degree of trapping produced by the electrode and the plate diameter. The highest Q is obtained by using mechanically or chemically polished blanks with an adequately large diameter and an evacuated enclosure.

In the vicinity of an isolated mode of vibration, the impedance of a crystal resonator is a pure resistance at two frequencies. The lower of these is the *resonance frequency* f_r (close to, but not exactly equal to the series self resonance frequency f_s due to the presence of C_0); the greater is the antiresonance frequency f_a. In the lossless case, the frequency of antiresonance is equal to the parallel resonance frequency f_p approximately equal to:

$$f_p = f_s \left(1 + \frac{C_1}{2C_0}\right) = f_s \left(1 + \frac{1}{2M}\right) \tag{9.167}$$

Table 9.3 Typical Specifications and Parameters for Precision SC-Cut Crystal Resonators

Parameter	*Specifications*	*Comments*
Frequency	5–160 MHz	Harmonic mode for higher frequencies
Recommended load capacitance	About 20 pF	Typically series capacitor for higher frequencies
Frequency adjustment tolerance	1.5–8 ppm	Generally, higher for higher frequency
Q	80k to 2.5M	Higher Q for lower frequency
R_1	35–120	
C_1 (fF)	0.13–0.5 fF	
C_0 (pF)	3.2–4.7 pF	

For resonators with a large figure of merit ($M > 5$), f_r can be approximated by:

$$f_r = f_s \left(1 + \frac{1}{2QM}\right) \tag{9.168}$$

Table 9.3 presents some typical parameters as found on product data sheets, for example, in [34].

The impedance of the crystal can be plotted as in Figure 9.98. It can be seen that the crystal is inductive in the region between ω_r (very close to ω_s) and ω_p and this will be a very narrow frequency range. If the crystal is used to replace an inductor in an oscillator circuit, for example, as shown in Figure 9.99(a), then oscillations will only occur in the frequency range where the crystal is actually inductive. While the crystal behaves like an inductor at the oscillating frequency, unlike a real inductor, no dc current flows through the crystal, because at dc, it is like a capacitor. Figure 9.99(b) can be derived by assuming the positive input in Figure 9.99(a) is grounded. This is the familiar Pearce amplifier, a subset of the Colpitts oscillator and is a very common way to construct a crystal oscillator. Figure 9.100 shows two ways to realize this Colpitts-based crystal oscillator: one with a bipolar transistor and one with MOS transistors.

As to the phase noise, besides the thermal noise with its floor around 160 dBc/Hz, $1/f$ noise exists in crystal oscillators. The total noise power spectral density of a crystal oscillator can be determined with Leeson's formula [8, 35].

$$\mathrm{PN} = \frac{|N_{\mathrm{OUT}}(s)|^2}{2P_S} = \frac{|H_1|\omega_o}{(2Q\ \omega)^{\div}}^2 \frac{|N_{\mathrm{IN}}(s)|^2}{2P_S}^{\div} \tag{9.169}$$

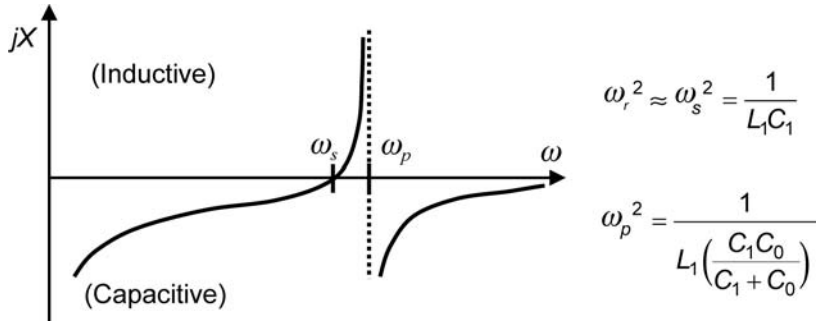

Figure 9.98 Impedance of crystal circuit model.

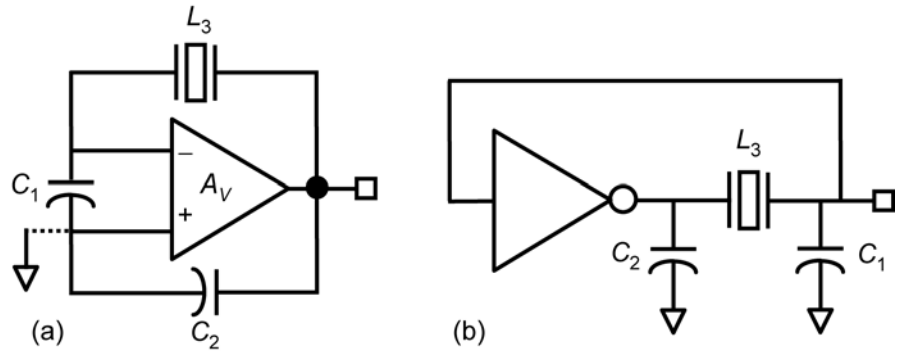

Figure 9.99 (a, b) Crystal oscillator diagrams using general amplifiers.

An empirical formula to describe crystal oscillator phase noise is given by:

$$\mathrm{PN}(f) = 10^{16\pm1} \times 1 + \left(\frac{f_0}{2f \times Q_L}\right)^2 \left(1 + \frac{f_c}{|f|}\right) \tag{9.170}$$

where f_0 is the oscillator output frequency, f is the offset frequency, and f_c is the corner frequency between 1/*f* and thermal noise regions, which is normally in the range 1 to 10 kHz. Q_L is the loaded Q of the resonator. Since the Q for a crystal resonator is very large as shown in Table 9.3, the reference noise contributes only to the very close-in noise and it quickly reaches the thermal noise floor at an offset frequency around f_c. Figure 9.101 demonstrates an example of a phase noise spectral density of a crystal reference source. This includes a plot of (9.170) using a first term of 10^{16}, an f_o of 10 MHz, Q_L of 120k, and f_c of 1 kHz, showing good agreement. Note that Q_L, the loaded Q, can be significantly lower than the unloaded Q, due to loading of the active circuitry

In summary, because of its effective high Q (up to hundreds of thousands) and relatively low frequency (less than a few hundred megahertz), crystal oscillators

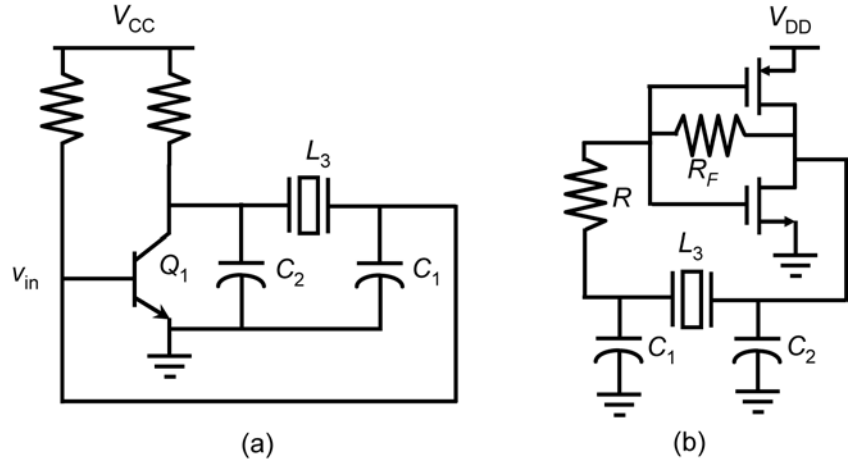

Figure 9.100 (a, b)Two implementations of crystal oscillators.

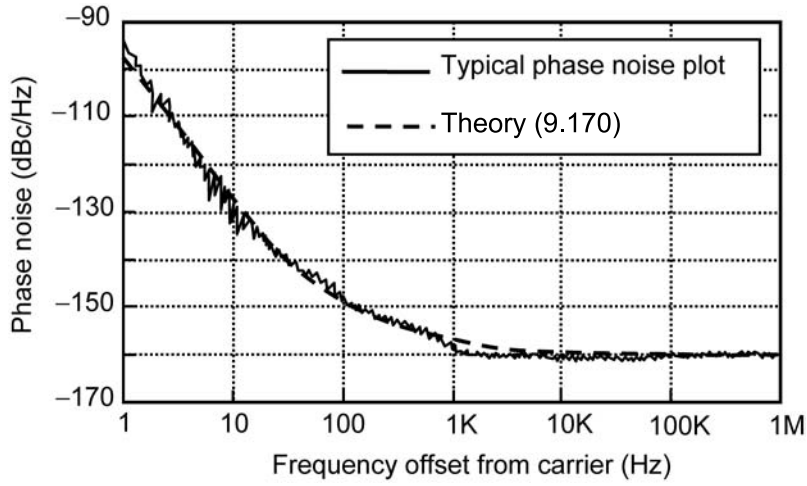

Figure 9.101 Phase noise of a crystal reference source. Theory with f_o = 10 MHz, Q_L = 120k, f_c = 1 kHz.

will have significantly lower phase noise and lower power dissipation than LC or ring oscillators. However, typically the purpose of a crystal oscillator is to achieve ultra-high stability, of the order of parts per million. To accomplish this, a complete commercial crystal oscillator will also have the means to do temperature compensation and amplitude control, and this is where a lot of the design effort would be directed.

References

[1] Kurokawa, K., "Some Basic Characteristics of Broadband Negative Resistance Oscillator Circuits," *The Bell System Technical Journal*, July 1969.

[2] Gonzalez, G., *Microwave Transistor Amplifiers Analysis and Design,* 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 1997.

[3] Voinigescu, S. P., D. Marchesan, and M. A. Copeland, "A Family of Monolithic Inductor-Varactor SiGe-HBT VCOs for 20 GHz to 30 GHz LMDS and Fiber-Optic Receiver Applications," *Proc. RFIC Symposium*, June 2000, pp. 173–176.

[4] Dauphinee, L., M. Copeland, and P. Schvan, "A Balanced 1.5 GHz Voltage Controlled Oscillator with an Integrated LC Resonator," *International Solid-State Circuits Conference*, 1997, pp. 390–391.

[5] Zannoth, M., et al., "A Fully Integrated VCO at 2 GHz," *IEEE J. Solid-State Circuits*, Vol. 33, December 1998, pp. 1987–1991.

[6] Hegazi, E., H. Sjoland, and A. Abidi, "A Filtering Technique to Lower LC Oscillator Phase Noise," *IEEE J. Solid-State Circuits,* Vol. 36, December 2001, pp. 1921–1930.

[7] Razavi, B., "A Study of Phase Noise in CMOS Oscillators," *IEEE J. Solid-State Circuits*, Vol. 31, March 1996, pp. 331–343.

[8] Leeson, D. B., "A Simple Model of Feedback Oscillator Noise Spectrum," *Proc. of the IEEE*, February 1966, pp. 329–330.

[9] Adams, S., *The Dilbert Principle*, New York: HarperCollins, 1996.

[10] Lee, T. H., and A. Hajimiri, "Oscillator Phase Noise: A Tutorial," *IEEE J. Solid-State Circuits*, Vol. 35, No. 3, March 2000, pp. 326–336.

[11] Mazzanti, A., and P. Andreani, "A 1.4mW 4.90-to-5.65GHz Class-C CMOS VCO with an Average FoM of 194.5dBc/Hz," *Proc. Int. Symp. Solid-State Circuits*, February 2008, pp. 474–475.

[12] Andreani, P., and S. Mattison, "On the Use of MOS Varactors in RF VCOs," *IEEE J. Solid-State Circuits,* Vol. 35, No. 6, June 2002, pp. 905–910.

[13] Fong, N., et al., "Accumulation MOS Varactors for 4 to 40 GHz VCOs in SOI CMOS," proc. *2002 International SOI Conference*, October 2002, pp. 158–160.

[14] Fong, N., et al., "A 1V 3.8-5.7 GHz Differentially-Tuned VCO in SOI CMOS," *Proc. 2002 RFIC Symp.*, June 2002, pp. 75–78.

[15] Fong, N., et al., "Phase Noise Improvement of Deep Submicron Low-Voltage VCO," *Proc. ESSCIRC* 2002, Florence, Italy, September 2002, pp. 811–814.

[16] Margarit, M., et al., "A Low-Noise, Low-Power VCO with Automatic Amplitude Control for Wireless Applications," *IEEE J. Solid-State Circuits,* Vol. 34, June 1999, pp. 761–771.

[17] Maneatis, J. G., "Low-Jitter Process-Independent DLL and PLL Based on Self-Biased Techniques," *IEEE J. Solid-State Circuits,* Vol. 31, No. 11, November 1996, pp. 1723–1732.

[18] Dai, L., and R. Harjani, "Design of Low-Phase-Noise CMOS Ring Oscillators," *IEEE Trans. Circuits and Systems II,* Vol. 49, No. 5, May 2002 pp. 328–338.

[19] Park, C. H., and B. Kim, "A Low-Noise, 900-MHz VCO in 0.6-m CMOS," *IEEE J. Solid-State Circuits,* Vol. 34, No. 5, May 1999, pp. 586–591.

[20] Eken, Y. A., and J. P. Uyemura, "A 5.9-GHz Voltage-Controlled Ring Oscillator in 0.18-μm CMOS," *IEEE J. Solid-State Circuits,* Vol. 39, No. 1, January 2004, pp. 230–233.

[21] Song, S. J., S. M. Park, and H. J. Yoo, "A 4-Gb/s CMOS Clock and Data Recovery Circuit Using 1/8-Rate Clock Technique," *IEEE J. Solid-State Circuits,* Vol. 38, No. 7, July 2003, pp. 1213–1219.

[22] Shu, Z., K. L. Lee, and B. H. Leung, "A 2.4-GHz Ring-Oscillator-Based CMOS Frequency Synthesizer with a Fractional Divider Dual-PLL Architecture," *IEEE J. Solid-State Circuits,* Vol. 39, No. 3, March 2004, pp. 452–462.

[23] Tang, J., D. Kasperkovitz, and A. Roermund, "A 9.8–11.5-GHz Quadrature Ring Oscillator for Optical Receivers," *IEEE J. Solid-State Circuits,* Vol. 37, No. 3, March 2002, pp. 438–442.

[24] Lee, S. J., B. Kim, and K. Lee, "A Novel High-Speed Ring Oscillator for Multiphase Clock Generation Using Negative Skewed Delay Scheme," *IEEE J. Solid-State Circuits,* Vol. 32, No. 2, February 1997, pp. 289–291.

[25] Eken, Y. A., and J. P. Uyenmura, "Multiple-GHz Ring and LC VCOs in 0.18μm CMOS," *Proc. RFIC*, Fort Worth, TX, June 2004, pp. 475–478.

[26] Hajimiri, A., S. Limotyrakis, and T. H. Lee, "Jitter and Phase Noise in Ring Oscillators," *IEEE J. Solid-State Circuits,* Vol. 34, No. 3, June 1999, pp. 790–804.

[27] McNeill, J. A., "Jitter in Ring Oscillators," *IEEE J. Solid-State Circuits,* Vol. 32, No. 6, June 1997, pp. 870–879.

[28] Gierkink, S. L. J., et al., "Intrinsic 1 Device Noise Reduction and Its Effect on Phase Noise in CMOS Ring Oscillators," *IEEE J. Solid-State Circuits,* Vol. 34, No. 7, July 1999, pp. 1022–1025.

[29] Thamsirianunt, M., and T. A. Kwasniewski, "CMOS VCO's for PLL Frequency Synthesis in GHz Digital Mobile Radio Communications," *IEEE J. Solid-State Circuits,* Vol. 32, No. 10, October 1997, pp. 1511–1524.

[30] Andreani, P., et al., "Analysis and Design of a 1.8 GHz CMOS LC Quadrature VCO," *IEEE J. Solid-State Circuits,* Vol. 37, No. 12, December 2002, pp. 1737–1747.

[31] van der Tang, J., et al., "Analysis and Design of an Optimally Coupled 5-GHz Quadrature LC Oscillator," *IEEE J. Solid-State Circuits,* Vol. 37, No. 5, May 2002, pp. 657–661.

[32] Gierkink, S. L., et al., "A Low-Phase-Noise 5-GHz CMOS Quadrature VCO Using Superharmonic Coupling," *IEEE J. Solid-State Circuits,* Vol. 37, No. 5, May 2002, pp. 1148–1154.

[33] Vig, J. R., *Quartz Crystal Resonators and Oscillators*, U.S. Army Electronics Technology and Devices Report, SLCET-TR-88-1, 1988.

[34] "Resonator Products," Piezo Technology Inc., Orlando, Florida, http://www.piezotech.com/Resonators/resonatorsindex.htm.

[35] Watanabe, Y., et al., "Phase Noise Measurements in Dual-Mode SC-Cut Crystal Oscillators," *IEEE Trans. on Ultrasonics, Ferroelectrics and Frequency Control*, Vol. 47, No. 2, March 2000, pp. 374–378.

Selected Bibliography

Chen, W., and J. Wu, "A 2-V 2-GHz BJT Variable Frequency Oscillator," *IEEE J. Solid-State Circuits*, Vol. 33, September 1998, pp. 1406–1410.

Craninckx, J., and M. S. J. Steyaert, "A 1.8-GHz Low-Phase-Noise CMOS VCO Using Optimized Hollow Spiral Inductors," *IEEE J. Solid-State Circuits,* Vol. 32, May 1997, pp. 736–744.

Craninckx, J., and M. S. J. Steyaert, "A Fully Integrated CMOS DCS-1800 Frequency Synthesizer," *IEEE J. Solid-State Circuits*, Vol. 33, December 1998, pp. 2054–2065.

Craninckx, J., and M. S. J. Steyaert, "A 1.8-GHz CMOS Low-Phase-Noise Voltage-Controlled Oscillator with Prescaler," *IEEE J. Solid-State Circuits*, Vol. 30, December 1995, pp. 1474–1482.

Hajimiri, A., and T. H. Lee, "A General Theory of Phase Noise in Electrical Oscillators," *IEEE J. Solid-State Circuits,* Vol. 33, June 1999, pp. 179–194.

Jansen, B., K. Negus, and D. Lee, "Silicon Bipolar VCO Family for 1.1 to 2.2 GHz with Fully-Integrated Tank and Tuning Circuits," *Proc. International Solid State Circuits Conference,* 1997, pp. 392–393.

Niknejad, A. M., J. L. Tham, and R. G. Meyer, "Fully-Integrated Low Phase Noise Bipolar Differential VCOs at 2.9 and 4.4 GHz," *Proc. European Solid-State Circuits Conference*, 1999, pp. 198–201.

Razavi, B., "A 1.8 GHz CMOS Voltage-Controlled Oscillator," *Proc. International Solid State Circuits Conference,* 1997, pp. 388–389.

Rogers, J. W. M., J. A. Macedo, and C. Plett, "The Effect of Varactor Nonlinearity on the Phase Noise of Completely Integrated VCOs," *IEEE J. Solid-State Circuits*, Vol. 35, September 2000, pp. 1360–1367.

Soyuer, M., et al., "A 2.4-GHz Silicon Bipolar Oscillator with Integrated Resonator," *IEEE J. Solid-State Circuits*, Vol. 31, February 1996, pp. 268–270.

Soyuer, M., et al., "An 11-GHz 3-V SiGe Voltage Controlled Oscillator with Integrated Resonator," *IEEE J. Solid-State Circuits*, Vol. 32, December 1997, pp. 1451–1454.

Svelto, F., S. Deantoni, and R. Castello, "A 1.3 GHz Low-Phase Noise Fully Tunable CMOS LC VCO," *IEEE J. Solid-State Circuits*, Vol. 35, March 2000, pp. 356–361.

CHAPTER 10

Frequency Synthesis

10.1 Introduction

There are many ways to realize synthesizers, but possibly the most common is based on a phase-locked loop (PLL). PLL-based synthesizers can be further subdivided by which type of a division is used in the feedback path. The division ratio N can be either an integer or a fractional number. If the number is fractional, then the synthesizer is called a fractional-N synthesizer. This type of synthesizer can be further distinguished by the method used to control the divide ratio, for example, by a sigma-delta controller or by some other technique. In this chapter, analysis is done with a general N without details of how N is implemented; thus, the analysis is applicable both to integer-N and fractional-N synthesizers. Additional general information on synthesizers can be found in [1–66].

10.2 Integer-*N* PLL Synthesizers

An integer-N PLL (phase-locked loop) is the simplest type of phase-locked loop synthesizer and is shown in Figure 10.1. Note that N refers to the divide-by-N block in the feedback of the PLL. The two choices are to divide by an integer, (integer N), or to divide by a fraction, (fractional N), essentially by switching between two or more integer values such that the effective divider ratio is a fraction. This is usually accomplished by using a modulator to control the division ratio. PLL-based synthesizers are among the most common ways to implement a synthesizer and this area is the subject of a great deal of research and development [20–52]. The PLL-based synthesizer is a feedback system that compares the phase of a reference f_r to the phase of a divided down output, f_o, of a controllable signal source (also known as a *voltage-controlled oscillator* or a VCO). The summing block in the feedback is commonly called a phase comparator or a phase detector. Through feedback, the loop forces the phase of the signal source to track the phase of the feedback signal and therefore their frequencies must be equal. Thus, the output frequency, which is a multiple of the feedback signal, is given by:

$$f_{\text{VCO}} = N \times f_r \tag{10.1}$$

Due to divider implementation details, it is not easily possible to make a divider that divides by noninteger values. Thus, a PLL synthesizer of this type is called an

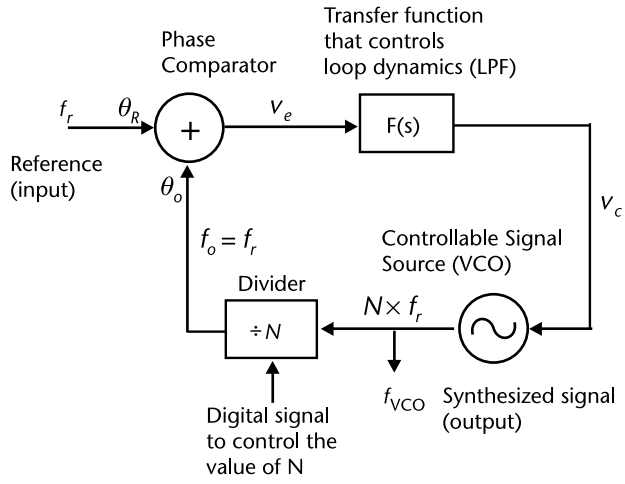

Figure 10.1 A simple integer-*N* frequency synthesizer.

integer-*N* frequency synthesizer. Circuits inside the feedback loop can be described by their transfer functions. These transfer functions can be designed to engineer the system dynamics to meet design specifications for the synthesizer. Typically, $F(s)$ is realized with a lowpass filter (LPF) or lowpass network. The details of the loop and the circuit components in the loop will be discussed later in this chapter.

In brief, the PLL is a feedback system that forces the divided-down VCO output phase to follow the reference signal phase. That is, it is a negative feedback loop with phases as the input and output signals. The loop is composed of a phase detector, a lowpass filter, a VCO, and a divider. The phase detector, which is the summing block of the feedback system, is used to compare output phase θ_o to reference phase θ_R. The lowpass filter is usually a linear transfer function that is placed in the system to control the loop's dynamic behavior including the settling time and transient response. The VCO generates the output signal and the divider divides the VCO output signal back down to the same frequency as the input. We use a feedback system based on phase rather than frequency because in any feedback loop without infinite dc gain there is always an error (finite error signal) between the input (reference) and the output. Thus, if we used a frequency-locked loop, then in most cases, there would be an error in the output frequency and it would not track the input as precisely as does a loop based on phase. The input reference in wireless communications is a quartz crystal. These crystals are low cost and can be made to resonate with extreme accuracy at a particular frequency determined by the physical properties of the crystal. Unfortunately, they can only resonate at frequencies as high as about 100 MHz and therefore cannot be used directly as an LO in RF applications. The other disadvantage to using a crystal directly is that there is no convenient way to tune their frequency. This is one of the main reasons that frequency synthesizers have become so popular. If the divider block is implemented using circuitry such that the divide ratio is programmable, then a range of frequencies can be obtained without the need to change the reference frequency.

Since N is an integer, the minimum step size of this synthesizer is equal to the reference frequency f_r. Therefore, in order to get a smaller step size, the reference

frequency must be made smaller. This is often undesirable so instead a fractional-*N* design is often used. This will be discussed later.

10.3 PLL Components

We will now briefly look at the basic components needed to make a PLL-based synthesizer and their basic governing equations. Only a very basic introduction to these components will be given in this chapter and only the most common forms of these circuits will be considered here.

10.3.1 Voltage Controlled Oscillators (VCOs) and Dividers

At the most basic level, all VCOs will have an output frequency with some dependence on the control voltage (or sometimes control current) as shown in Figure 10.2. Note that the curve is not always linear (actually it is hardly ever linear, but for the moment we will assume that it is). Also, note that the control voltage can usually be adjusted between ground and the power supply voltage and that, over that range, the VCO frequency will move between some minimum and some maximum values.

Here, V_{C_nom} is the nominal control voltage coming from the loop filter, and ω_{nom} is the nominal frequency at this nominal voltage. Usually when considering loop dynamics, we only consider frequency deviations away from the nominal frequency ω_{VCO}, and voltage deviations away from the nominal voltage v_c. Thus, we can write the oscillating frequency as:

$$\omega_o = \omega_{nom} + \omega_{VCO} = \omega_{nom} + K_{VCO} v_c \tag{10.2}$$

where:

$$v_c = v_C \quad V_{C\_nom} \tag{10.3}$$

In addition, if we remove this "dc" part of the equation and only consider the part that is changing, we are left with:

$$\omega_{VCO} = K_{VCO} v_c \tag{10.4}$$

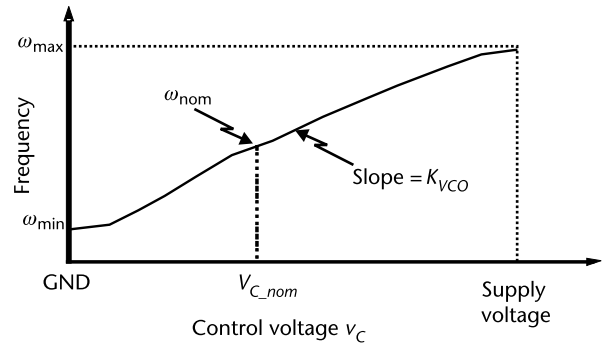

Figure 10.2 A typical VCO characteristic.

However, we would like to have an expression relating input voltage to output phase since the output of the VCO ultimately goes to the phase detector. To relate frequency ω to phase θ, we note that:

$$\omega = \frac{d\theta}{dt} \tag{10.5}$$

Therefore, the output phase of the VCO can be given as:

$$\theta_{\text{VCO}} = \omega_{\text{VCO}}\, dt = K_{\text{VCO}} \int_0^t v_c(\tau) d\tau \tag{10.6}$$

In the Laplace domain, this becomes:

$$\frac{\theta_{\text{VCO}}(s)}{v_c(s)} = \frac{K_{\text{VCO}}}{s} \tag{10.7}$$

Thus, we have the desired equation for the transfer function of the VCO block. Note that, for the purposes of system behavior, the divider can be thought of as an extension of the VCO. The output phase after the divider is simply:

$$\frac{\theta_o}{v_c} = \frac{1}{N} \times \frac{K_{\text{VCO}}}{s} \tag{10.8}$$

10.3.2 Phase Detectors

A phase detector produces an output signal proportional to the phase difference of the signals applied to its inputs. The inputs and outputs can be sine waves, square waves, or other periodic signals, not necessarily having a 50% duty cycle. The output signal could be a current or voltage and it could have multiple frequency components. Since the dc value is the component of interest, the phase detector is typically followed by some sort of filter. Thus, the equation that describes a phase detector is:

$$v_e(s) = K_{\text{phase}}(\theta_R(s) \quad \theta_o(s)) \tag{10.9}$$

provided that the phase detector is a continuous time circuit (which is often not the case in integrated circuit implementations). The output of the phase detector, $v_e(s)$, is often also called the error voltage and is seen to be proportional to the difference of the input phases with proportionality constant K_{phase}. This is a linearized equation, often valid only over limited range. Another comment that can be made about phase detectors is that often they respond in some way to a frequency difference as well. In such a case, the circuit is often referred to as a *phase-frequency detector* (PFD).

The phase detector can be as simple as an XOR gate or an XNOR gate, but typical phase detectors are usually more complicated. For example, flip-flops form the basis of tri-state phase detectors which is often combined with a charge pump. This type of phase detector has two outputs as shown in Figure 10.3. If the refer-

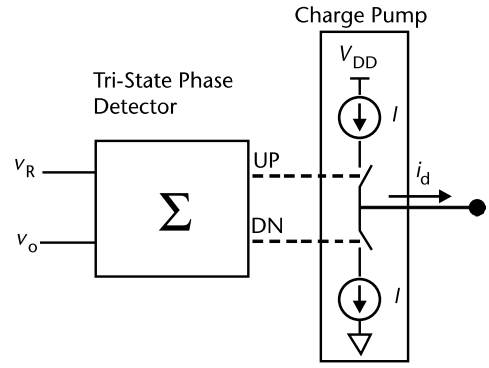

Figure 10.3 Tri-state phase detector and charge pump.

ence phase (v_R) is ahead of the output phase (v_o), then the circuit produces a signal UP that tells the VCO to speed up and therefore advance its phase to catch up with the reference phase. Conversely, if the reference phase is lagging the output phase, it produces a signal DN that tells the VCO to slow down and therefore retard its phase to match the reference phase. If the reference and output are in phase, then the phase detector does not produce an output.

The two digital signals produced by a PFD have to be converted back into an analog control signal at the input of the VCO, and the circuit most commonly used to do this is called a charge pump. A charge pump is made of two controllable current sources connected to a common output also shown in Figure 10.3. The outputs from the phase detector turn on one of the two currents, which either charge or discharge capacitors attached to the VCO input.

A simple implantation of the PFD is shown in Figure 10.4(a), and a description of the PFDs operation based on the state diagram is shown in Figure 10.4(b). Transitions happen only on the rising edge of v_o or v_R. Let us assume we start in the

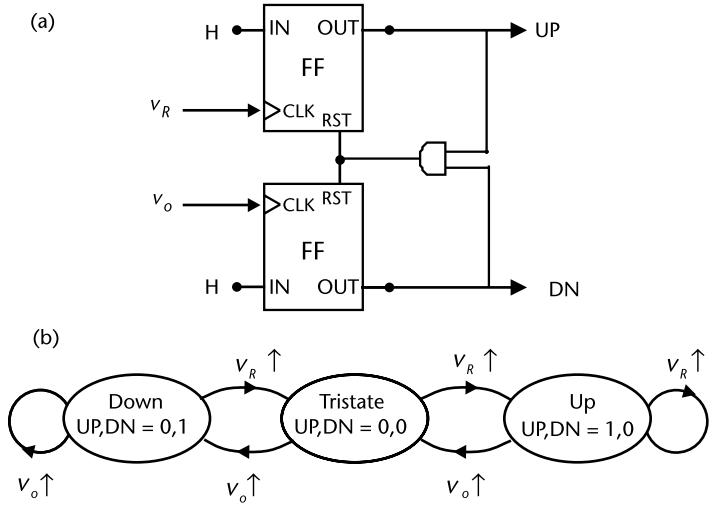

Figure 10.4 (a) Simple PFD implementation and (b) PFD state diagram.

middle state, the tri-state where both outputs are zero. Then, depending on which edge arrives first, the PFD moves either to the up state or to the down state. If the reference edge arrives first, the output needs to catch up and so the up switch turns on to charge up the output. It stays up until the other edge comes along; thus, the average output current from the charge pump depends on how far apart the two signals are. On the other hand, if the reference is behind the output, then the output is too fast and needs to be slowed down. This causes a down pulse to be generated, and as a result, current flows out of the charge pump, discharging the capacitors. The current flows for the length of time τ between the output edge and the reference edge. If the period is T, the average output current is:

$$i_d = I\frac{\tau}{T} = \frac{I}{2\pi}\div(\theta_R \quad \theta_o) \tag{10.10}$$

Thus K_{phase} for this phase detector is:

$$K_{\text{phase}} = \frac{I}{2\pi} \tag{10.11}$$

where I is the current that flows through the controllable current sources in the charge pump when they are on.

The operation of the PFD and current sources is shown in Figure 10.5. The movement from state to state is controlled by the rising edge only, so the pulse width of the input signals is not important. We have shown the pulses as being narrow, but it would work equally well using wider pulses. In Figure 10.5 it is shown that for the same phase difference between v_o and v_R, the output current depends on where the operation starts. In Figure 10.5(a), the v_o edge comes first, resulting in down pulses and a negative average output current. In Figure 10.5(b), the v_R edge comes first, resulting in up pulses and a positive average output current. We note also that if the v_o pulse was delayed (moved towards the right) in Figure 10.5(a), the down pulses would become narrower resulting in average current closer to zero. In Figure 10.5(b), for the same delay of the v_o pulses, the up pulses become wider and the average current moves closer to I. Note that the short pulses associated with UP in Figure 10.5(a) and DN in Figure 10.5(b) are realistic. They are the result of details of the PFD design.

If this average output current is now plotted as a function of the phase difference, with v_R taken as the reference, the result can be interpreted as the transfer function of the phase detector and charge pump and is shown in Figure 10.6. We

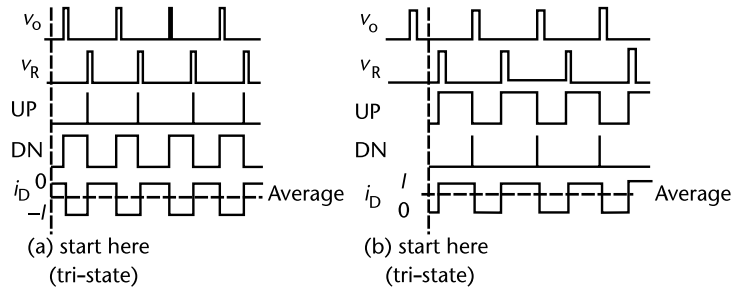

Figure 10.5 (a, b) Operation of PFD and current sources.

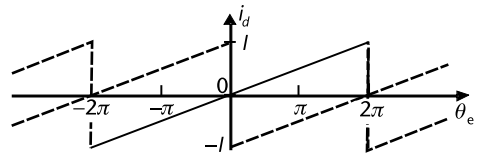

Figure 10.6 Average output current versus phase for PFD and charge pump.

note that any positive phase, for example, 60°, for which the output current would be I 60/360, could be interpreted as the equivalent negative phase, for example, +60° is equivalent to 300°, for which the current is equal to I 300/360. Thus, for every phase difference, there are two possible interpretations shown by the solid and dashed lines in Figure 10.6. We note that this is equivalent to starting the phase detector in a different state or at a different time, as was shown in Figure 10.5.

To illustrate how this phase detector can also be used as a frequency detector, in Figure 10.7, waveforms are shown for two different input frequencies. We have assumed that we start in tri-state. Since the output pulse v_o occurs first, down pulses DN occur, which would result in negative output current pulses i_d and an average negative output current, shown by the dotted line. However, since the frequencies are different, the output pulse width is changing, in this case becoming narrower and the average current is moving towards zero. Eventually, the phase detector experiences a second reference pulse v_R before the output pulse v_o and moves into the up state, and up current pulses i_d result. From then on, the phase detector output states will be either tri-state or in the up state, so only a positive current is ever provided. In this way, it can be seen that for a reference frequency higher than the output frequency, the average current is positive. Similarly, for a reference frequency lower than the output frequency, the average output current would always be negative (except, of course, for a possible short time at startup). Thus, with the correct feedback polarity, this current can be used to correct the VCO frequency until it is the same as the reference frequency and the loop is locked.

10.3.3 The Loop Filter

Normally, VCOs are controlled by voltage and not current. Thus, generally we need a method to turn the current produced by the charge pump back into a voltage.

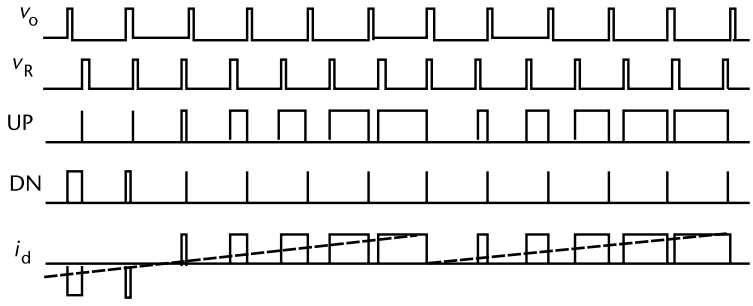

Figure 10.7 Output pulses for inputs at different frequencies.

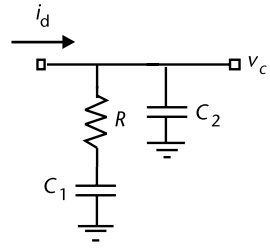

Figure 10.8 A typical loop filter.

In addition, lowpass filtering is needed since it is not desirable to feed pulses into the VCO. This is usually done by dumping the charge produced by the charge pump onto the terminals of a capacitor. As we will show later, a simple capacitor all by itself does not yield a stable loop, so a combination of capacitors and resistors are used. This part of the PLL is typically called the loop filter. One of the most common loop filters used is shown in Figure 10.8. Note that PLLs that do not use charge pumps have loop filters as well. In addition to turning the current back into the voltage, loop filters are also the components most commonly used to control system-level loop dynamics.

The frequency response of the phase-frequency detector, charge pump, and loop filter is mainly determined by the loop filter. The frequency response of the network (seen in Figure 10.9) will now be analyzed, starting with the admittance of the capacitor and resistor circuit.

$$Y = sC_2 + \frac{1}{R + \dfrac{1}{sC_1}} = sC_2 + \frac{sC_1}{sC_1R + 1} = \frac{sC_2(sC_1R + 1) + sC_1}{sC_1R + 1} \tag{10.12}$$

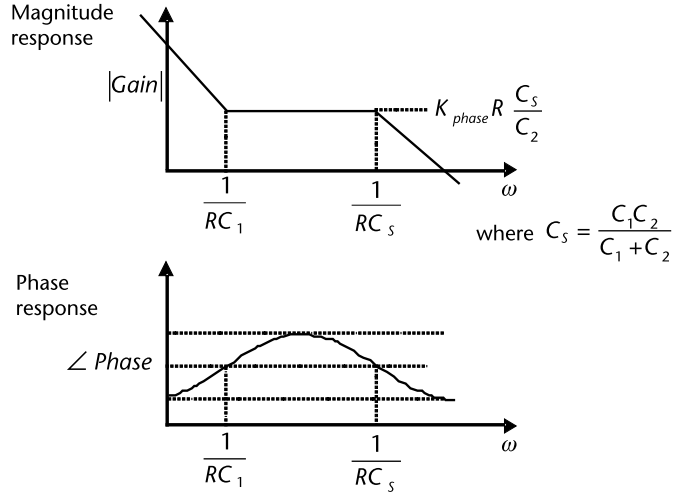

Figure 10.9 Phase-frequency detector, charge pump, and loop filter frequency response.

This admittance can be used to determine v_c, the control voltage of the VCO using (10.11):

$$v_c = \frac{i_d}{Y} = \frac{K_{\text{phase}}(\theta_R \quad \theta_o)(sC_1R+1)}{sC_2(sC_1R+1)+sC_1} = \frac{K_{\text{phase}}(\theta_R \quad \theta_o)(1+sC_1R)}{s(C_1+C_2)(1+sC_sR)} \tag{10.13}$$

where $C_s = \dfrac{C_1C_2}{C_1+C_2}$ and $K_{\text{phase}} = \dfrac{I}{2\pi}$.

The frequency response of the charge pump and filter as given in (10.13) is shown in Figure 10.9. We note that at low frequencies, the response is dominated by the pole at the origin in the transfer function and thus the circuit acts like an integrator. Also note that this has been derived in continuous time and, as long as the pulses are much faster than any changes of interest at the output, this is a reasonable assumption.

10.4 Continuous-Time Analysis for PLL Synthesizers

The s domain model for a synthesizer is shown in its most general form in Figure 10.10. Here, any loop filter (in combination with a charge pump) is simply shown as $F(s)$, and the dc gain is brought out explicitly as term A_o. We can therefore derive the basic loop transfer function for the loop.

10.4.1 Simplified Loop Equations

The overall transfer function is:

$$\frac{\theta_o}{\theta_R} = \frac{\dfrac{A_oK_{\text{phase}}F(s)}{N}\times\dfrac{K_{\text{VCO}}}{s}}{1+\dfrac{A_oK_{\text{phase}}F(s)}{N}\times\dfrac{K_{\text{VCO}}}{s}} = \frac{KF(s)}{s+KF(s)} \tag{10.14}$$

where K is given by:

$$K = \frac{A_oK_{\text{phase}}K_{\text{VCO}}}{N} \tag{10.15}$$

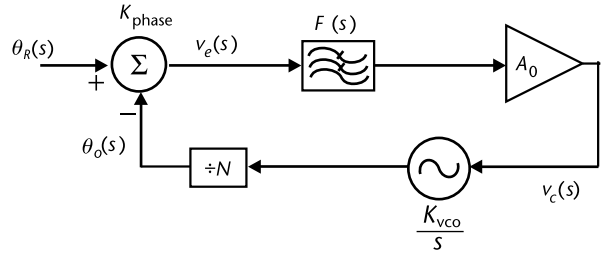

Figure 10.10 Complete loop in the frequency domain.

Now (10.14) is the most general PLL loop equation, and for specific loops it differs only in the form that $F(s)$ takes. For instance, in a first-order loop $F(s)$ is simply equal to 1. In this case, the loop equation becomes:

$$\frac{\theta_o}{\theta_R} = \frac{K}{s+K} \tag{10.16}$$

Note that for this first-order loop, for a phase ramp (change in frequency), the phase error is not zero because there are not enough integrators in the loop. Since zero phase error is often highly desired and due to a lack of flexibility, this loop is not often used in practice.

A much more common PLL is the second-order PLL. A typical second-order PLL has a loop filter with a transfer function of:

$$F(s) = \frac{\tau s + 1}{s} \tag{10.17}$$

A charge pump and PFD-based PLL with the loop filter, as previously discussed, is an example of a third-order loop, but as long as C_2 can be ignored, it can be treated as a second-order loop, as will now be shown. The most common system level configuration is shown in Figure 10.11. In this case (assuming for the moment that we ignore C_2), the impedance of the loop filter is:

$$F(s) = R + \frac{1}{sC_1} = \frac{sC_1R + 1}{sC_1} \tag{10.18}$$

Note this transfer function is an impedance since this stage converts current to voltage, which is not a typical transfer function, but works here. Thus, we can substitute this back into (10.14) and therefore find:

$$\frac{\theta_o}{\theta_R} = \frac{\dfrac{IK_{VCO}}{2\pi \times N}\left(R + \dfrac{1}{sC_1}\div\right)}{s + \dfrac{IK_{VCO}}{2\pi \times N}\left(R + \dfrac{1}{sC_1}\div\right)} = \frac{\dfrac{IK_{VCO}}{2\pi \times NC_1}(RC_1s + 1)}{s^2 + \dfrac{IK_{VCO}}{2\pi \times N}Rs + \dfrac{IK_{VCO}}{2\pi \times NC_1}} \tag{10.19}$$

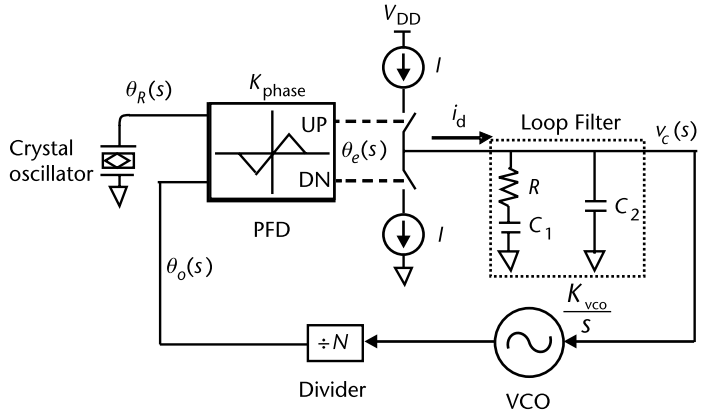

Figure 10.11 A frequency synthesizer implemented with a charge pump and PFD.

Thus, for this PLL, we get a second-order transfer function with a zero. Note the purpose of R can be seen directly from this equation. If R is set equal to zero, it can be seen by an inspection of (10.19) that the poles of this equation will sit on the $j\omega$ axis and the loop would oscillate or be on the verge of oscillating. From (10.19), expressions for the loop dynamics can be determined. The natural frequency of the loop is given by:

$$\omega_n = \sqrt{\frac{IK_{\text{VCO}}}{2\pi \times NC_1}} \tag{10.20}$$

The damping constant is given by:

$$\zeta = \frac{R}{2}\sqrt{\frac{IK_{\text{VCO}}C_1}{2\pi \times N}} \tag{10.21}$$

Often the resistor and capacitor value are to be determined for a known damping constant and natural frequency. It is straightforward to solve these two equations for these variables:

$$C_1 = \frac{IK_{\text{VCO}}}{2\pi \times N\omega_n{}^2} \tag{10.22}$$

and:

$$R = 2\zeta\sqrt{\frac{2\pi \times N}{IK_{\text{VCO}}C_1}} = \zeta\frac{4\pi \times N\omega_n}{IK_{\text{VCO}}} \tag{10.23}$$

From the above it can be shown that (10.19) can be rewritten in a general form as:

$$\frac{\theta_{\text{o}}}{\theta_R} = \frac{\omega_n{}^2\left(\frac{2\zeta}{\omega_n}s + 1\right)}{s^2 + 2\zeta\omega_n s + \omega_n{}^2} \tag{10.24}$$

This shows that there is a relationship between the pole and zero locations.

Note that it is easy to determine the transfer function even if the output is taken from other places in the loop. For instance, it is often interesting to look at the control voltage going into the VCO. In this case, the system transfer function becomes:

$$\frac{v_{\text{C}}}{\theta_R} = \frac{\frac{I \times s}{2\pi \times C_1}(RC_1 s + 1)}{s^2 + \frac{IK_{\text{VCO}}}{2\pi \times N}Rs + \frac{IK_{\text{VCO}}}{2\pi \times NC_1}} = \frac{\frac{N\omega_n{}^2}{K_{\text{VCO}}}s\left(\frac{2\zeta}{\omega_n}s + 1\right)}{s^2 + 2\zeta\omega_n s + \omega_n{}^2} \tag{10.25}$$

This expression contains an extra s in the numerator. This makes sense because the control voltage is proportional to the frequency of the VCO, which is the derivative of the phase of the output. Note that we can also write an expression for

the output frequency (as given by the control voltage) as a function of the input frequency, noting that frequency is the derivative of phase and starting from (10.25):

$$\frac{V_C}{\theta_R s} = \frac{\dfrac{I\times s}{2\pi\times C_1}(RC_1 s+1)}{s^2+\dfrac{IK_{VCO}}{2\pi\times N}Rs+\dfrac{IK_{VCO}}{2\pi\times NC_1}}\ \frac{1}{s} = \frac{\dfrac{N\omega_n{}^2}{K_{VCO}}s\ \dfrac{2\zeta}{\omega_n}s+1\div}{s^2+2\zeta\omega_n s+\omega_n{}^2}\ \frac{1}{s} \tag{10.26}$$

$$\frac{V_C}{\omega_R} = \frac{\dfrac{N\omega_n{}^2}{K_{VCO}}\ \dfrac{2\zeta}{\omega_n}s+1\div}{s^2+2\zeta\omega_n s+\omega_n{}^2}$$

which is nearly identical to the expression for phase transfer function θ_o/θ_R given by (10.24).

10.4.2 PLL System Frequency Response and Bandwidth

Figure 10.12 is a plot of the closed loop transfer function for the charge pump and PFD-based PLL, which is described by (10.19) and (10.24) for different values of damping constant. This diagram shows that the loop's 3-dB bandwidth is highly dependent on the damping constant. It can be shown that the 3-dB bandwidth of this system is given by [2]:

$$\omega_{3dB} = \omega_n\sqrt{1+2\zeta^2+\sqrt{4\zeta^4+4\zeta^2+2}} \tag{10.27}$$

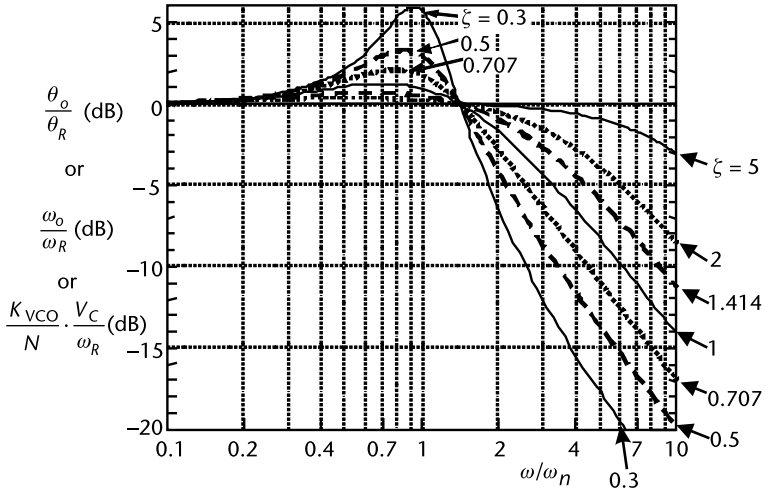

Figure 10.12 PLL frequency response of the closed loop transfer function of a high-gain second-order loop. Note that the graph is valid for any of the three functions shown on the *y*-axis.

Since this equation can be tedious without a calculator, two equations sometimes used to approximate this are:

$$\begin{aligned} &\omega_{3\text{dB}} \quad 2\zeta\omega_n \quad \zeta > 1.5 \quad \text{(approximation \#1)} \\ &\omega_{3\text{dB}} \quad (1+\zeta\sqrt{2})\omega_n \quad \zeta < 1.5 \quad \text{(approximation \#2)} \end{aligned} \tag{10.28}$$

10.4.3 Complete Loop Transfer Function Including C_2

Note that if C_2 is included this adds a high frequency pole to the system. Normally this capacitor is chosen to be about one-tenth of the value of C_1 and is included to clean up high-frequency ripple on the control line. If this capacitor is included, then the following expression for open-loop gain can be derived:

$$\left.\frac{\theta_o}{\theta_R}\right|_{\text{open loop}} = \frac{K_{\text{VCO}}K_{\text{phase}}(1+sC_1R)}{s^2N(C_1+C_2)(1+sC_sR)} \tag{10.29}$$

Here C_s is the series combination of C_1 and C_2. This is plotted in Figure 10.13, which shows a low-frequency slope of –40 dB/decade and 180° of phase shift. After the zero, the slope is –20 dB per decade and the phase heads back towards 90° of phase shift. After the high-frequency pole, the slope again is –40 dB/decade and the phase approaches 180°. Note that the dashed lines in the graph show the response of the system if the capacitor C_2 is not included. For optimal stability (maximum phase margin in the system), the unity gain point should be at the geometric mean of the zero and the high-frequency pole, since this is the location where the phase shift is furthest from 180°. Some may wonder, after seeing this plot, if the system is actually unstable at dc because, at low frequencies, the phase shift is 180° and

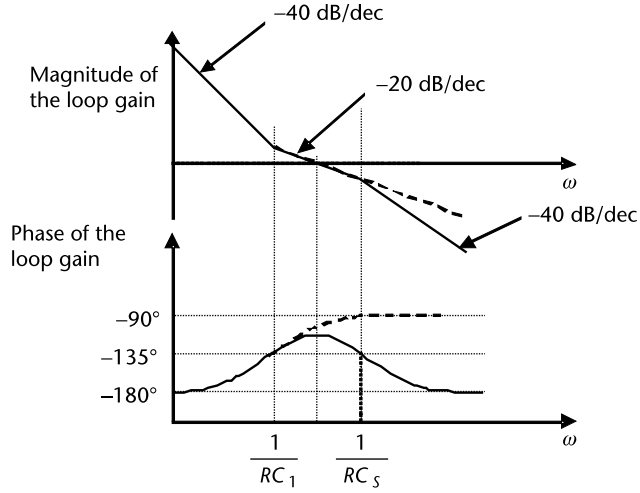

Figure 10.13 Open-loop magnitude and phase response. Note that the dotted line shows a response if a high-frequency pole is not included.

the gain is greater than 1. In fact, this is not the case and the system is stable. A full stability analysis, for example, by plotting the closed-loop poles, would show this.

The closed-loop gain with C_2 is given by:

$$\frac{\theta_o}{\theta_R} = \frac{K_{\text{VCO}}K_{\text{phase}}(1 + sC_1R)}{s^2N(C_1 + C_2)(1 + sC_sR) + K_{\text{VCO}}K_{\text{phase}}(1 + sC_1R)} \tag{10.30}$$

Thus, one could now estimate all the parameters of the loop. Figure 10.13 shows that if the zero and the high-frequency pole are relatively far apart, then up to the unity gain point, the loop parameters are nearly the same whether or not the high-frequency pole is included. There is, however, a slight decrease of phase margin (in the diagram from about 75° to about 65°).

A cautionary note about the choice of C_2 should also be mentioned now. It may not always be appropriate to use a value of one-tenth C_1 for the value of C_2. It has been assumed in this section that most of the time C_2 does not significantly change the loop dynamics; however, R increases at high ζ, and in this case the impedance of the series combination of C_1 and R may become comparable to the impedance of C_2 at the loop natural frequency. If this happens, then the equations derived in this section will become increasingly inaccurate, in which case there would be no choice but to reduce the value of C_2 to less than $C_1/10$.

Note that while in practice C_2 is almost always present, the results in the previous sections can be applied almost always to a third-order loop with a very small error. Thus, these results are also valid for this more practical case.

10.5 Discrete Time Analysis for PLL Synthesizers

The preceding linear continuous time analysis of the synthesizer is not valid under all conditions. If the loop bandwidth is increased so that it becomes a significant fraction of the reference frequency, the previous analysis becomes increasingly inaccurate. For this reason, it is sometimes necessary to treat the synthesizer system as a discrete time control system as it truly is [19]. To do this, we must consider the PFD, which in this case is the sampling element. We assume that in lock the PFD produces only narrow impulses at the reference frequency. Note that because the charge pump behaves like an integrator, it has infinite gain at dc. Thus, as long as the frequency deviation is small, this high gain drives the phase error towards zero and the output pulses will be narrow. Therefore, it acts like an ideal sampler. The loop filter holds the charge dumped onto it in each cycle, so, in this system, it acts as a hold function. Hence, this is a sampled system, and the s-domain combination of the VCO, divider, PFD, and loop filter must be converted into their z-domain equivalents (being careful to remember to multiply the functions together before converting them into the z-domain). Thus, the closed-loop transfer function, including the sampling action of these four blocks, is (ignoring C_2) (see Figure 10.14) [2]:

$$G(z) = \frac{K(z \quad \alpha)}{z^2 + (K \quad 2)z + (1 \quad \alpha K)} \tag{10.31}$$

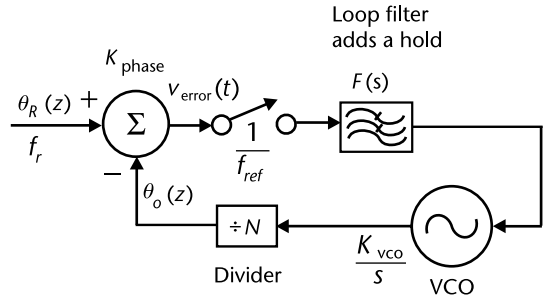

Figure 10.14 Discrete-time system-level diagram for a synthesizer.

where

$$\alpha = \frac{4\zeta \quad \omega_n T}{4\zeta + \omega_n T} \tag{10.32}$$

$$K = \frac{\omega_n{}^2 T^2}{2} \quad 1 + \frac{4\zeta}{\omega_n T} \div \tag{10.33}$$

and T is the period of the reference.

Starting from the closed loop transfer function, the pole locations as a function of the reference period T can be sketched and are shown in Figure 10.15. Note that depending on the specific parameters, this plot will change slightly, but the basic shape of the root locus will remain the same. The point of greatest interest here is that point at which the reference period is increased to some critical value and the system becomes unstable, as one of the poles moves outside the unit circle. At or even close to this period the *s*-domain analysis discussed in the previous section will be increasingly inaccurate. Note that the reference frequency would not normally be this low in the design of a PLL, but the designer must be aware of this so that the assumptions of the previous section are not violated.

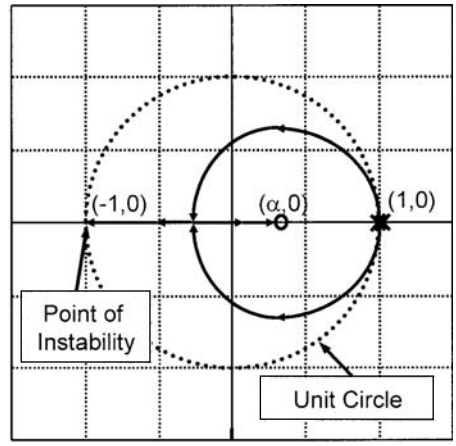

Figure 10.15 Sketch of closed-loop pole location as a function of increasing reference period.

Now the poles of (10.31) are given by:

$$\text{Poles} = 1 \quad \frac{K}{2} \pm \frac{1}{2}\sqrt{(K \quad 2)^2 \quad 4(1 \quad \alpha K)} \tag{10.34}$$

The pole that has the larger positive value is not of concern, because it will never leave the unit circle. However, it is of interest to find out when:

$$1 \quad \frac{K}{2} \quad \frac{1}{2}\sqrt{(K \quad 2)^2 \quad 4(1 \quad \alpha K)} = \quad 1 \tag{10.35}$$

Skipping a number of steps (primarily because they are boring), this will happen when:

$$K(1+\alpha) = 4 \tag{10.36}$$

Taking this expression and substituting back in for K and α, the critical period for which the loop will go unstable, T_{US}, is:

$$T_{US} = \frac{1}{\omega_n \zeta} \tag{10.37}$$

Noting that:

$$T_{US} = \frac{2\pi}{\omega_{\text{ref\_crt}}} \tag{10.38}$$

where $\omega_{\text{ref\_crt}}$ is the reference frequency at which the loop goes unstable. It can be determined that:

$$\omega_{\text{ref\_crt}} = 2\pi\zeta\omega_n \tag{10.39}$$

Therefore:

$$\frac{\omega_{\text{ref}}}{\omega_n} \quad 2\pi\zeta \tag{10.40}$$

So, for instance, in the case of $\zeta = 0.707$, this ratio must be greater than 4.4. Therefore, for a reference frequency of 40 MHz, if the loop natural frequency is set any higher than 9.1 MHz, the loop will go unstable. A number that is quoted often as being a "safe" ratio is 10 to 1 [15].

10.6 Transient Behavior of PLLs

In the two previous sections, linear *s*-domain equations that describe the PLL as a classic feedback system and more complicated *z*-domain equations were derived. However, the behavior of a real PLL is much more complex than either of these two analyses can explain. This is because until the loop starts tracking the phase of the input, or alternatively if there is a very large step or ramp in the phase of the input, the loop's output phase may not be able to follow the input phase. This is primarily due to the limitations of the phase detector, which has a narrow linear range. For

example, the tri-state PDF has a linear range of $\pm 2\pi$. If an event at the input occurs that causes the phase error to exceed 2π, then the loop will experience a nonlinear event: cycle slipping. Remember that in the previous analysis it was assumed that the phase detector was linear. This nonlinear event will cause a transient response that cannot be predicted by the theory of the previous section. The loop will of course work to correct the phase and force the VCO to track the input once more. When the loop goes into this process, it is said to be in acquisition mode as it is trying to acquire phase lock, but has not yet acquired it. Note that acquisition also happens when the PLL is first powered, since the VCO and reference will be at a random phase and probably will have a frequency difference. In extreme cases of this, the VCO may even be forced beyond its linear range of operation, which may result in the loop losing lock indefinitely. These situations will now be explored. First, the case where the loop is in lock and experiences no cycle slipping will be considered.

10.6.1 PLL Linear Transient Behavior

Here, the linear transient response of the most common PLL presented in this chapter will be considered further. In this section, only the *s*-domain response will be discussed, which under most normal operating conditions is sufficient. However, the *z*-domain equivalent of this analysis could be undertaken. For the linear transient behavior, the phase error rather than the output phase is needed, so a different transfer function has to be derived for the system shown in Figure 10.10. The result is:

$$\frac{\theta_e}{\theta_R} = \frac{s^2}{s^2 + 2\zeta\omega_n s + \omega_n{}^2} \tag{10.41}$$

In this section, we will see the response to an input frequency step ω. Since the input is described by phase, we take the phase equivalent of a frequency step, which is equivalent to a ramp of phase (we note that phase is the integral of frequency and the integral of a step is a ramp). Thus, the input is described by:

$$\theta_R = \frac{\omega}{s^2} \tag{10.42}$$

This input in (10.42) when multiplied by the transfer function (10.41) results in:

$$\theta_e = \frac{\omega}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{10.43}$$

Then, the inverse Laplace transform is taken, with the following results:

$$\theta_e(t) = \frac{\omega}{\omega_n} \frac{\sinh \omega_n \sqrt{\zeta^2 \quad 1}\, t}{\sqrt{\zeta^2 \quad 1}} e^{\ \zeta \omega_n t} \quad \zeta > 1 \tag{10.44}$$

$$\theta_e(t) = \frac{\omega}{\omega_n}\omega_n t \times e^{\omega_n t} \quad \zeta = 1 \tag{10.45}$$

$$\theta_e(t) = \frac{\omega}{\omega_n}\frac{\sin\omega_n\sqrt{1-\zeta^2}\,t}{\sqrt{1-\zeta^2}}\, e^{\zeta\omega_n t} \quad \zeta < 1 \tag{10.46}$$

These results are plotted in Figure 10.16 for various values of damping constant.

It can be seen that a damping constant of 0.707 to 1 results in the fastest settling (reduction of phase error to zero). Depending on the required level of settling, one can determine the settling time. To the accuracy of the above diagram, settling is better than 99% complete when $\omega_n t = 7$ for $\zeta = 0.707$. Thus, given a required settling time, one can calculate the required natural frequency. To prevent the reference frequency from feeding through to the VCO, the loop bandwidth, as shown in Figure 10.12 and estimated in (10.28), must be significantly less than the reference frequency. In fact, the extra capacitor in the loop filter has been added in order to provide attenuation at the reference frequency.

It is also interesting to look at the control voltage:

$$\frac{V_C}{\omega_R} = \frac{\frac{N\omega_n^2}{K_{\text{VCO}}}\left(\frac{2\zeta}{\omega_n}s + 1\right)}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{10.47}$$

In this case again we apply a step in frequency so that:

$$\omega_R = \frac{\omega}{s} \tag{10.48}$$

Note that this equation is given in the frequency rather than the phase domain, so s is raised to unity power in the denominator. Therefore, the control voltage is

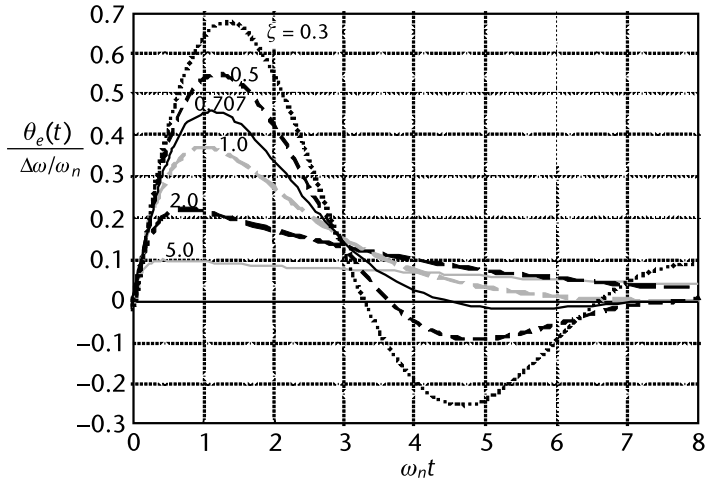

Figure 10.16 Error for frequency step, high-gain second-order loop. Note $\theta_e = \theta_R - \theta_o$.

given by:

$$V_C = \frac{\frac{N\omega_n^2}{K_{VCO}}\left(\frac{2\zeta}{\omega_n}s + 1\right)}{s^2 + 2\zeta\omega_n s + \omega_n^2} \times \frac{\omega}{s} = \frac{\frac{2\zeta \times N\omega_n}{K_{VCO}} \times \omega}{s^2 + 2\zeta\omega_n s + \omega_n^2} + \frac{\frac{N\omega_n^2}{K_{VCO}} \times \frac{\omega}{s}}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{10.49}$$

Now the first term is simply a scaled version of the previous expression, and the second term is the integral of the first term. Therefore, the transient expression for the control voltage is given by:

$$V_C(t) = \frac{\zeta \times N \quad \omega}{K_{VCO}} \quad \frac{\sinh \omega_n \sqrt{\zeta^2 \quad 1}t}{\sqrt{\zeta^2 \quad 1}} \quad e^{\ \zeta\omega_n t} \quad \frac{N \quad \omega}{K_{VCO}} \cosh \omega_n \sqrt{\zeta^2 \quad 1}t \quad e^{\ \zeta\omega_n t} \tag{10.50}$$

$$+ \frac{N \quad \omega}{K_{VCO}} \quad \zeta > 1$$

$$V_C(t) = \frac{N \quad \omega}{K_{VCO}} \omega_n t \times e^{\ \omega_n t} \quad \frac{N \quad \omega}{K_{VCO}} \times e^{\ \omega_n t} + \frac{N \quad \omega}{K_{VCO}} \quad \zeta = 1 \tag{10.51}$$

$$V_C(t) = \frac{\zeta \times N \quad \omega}{K_{VCO}} \quad \frac{\sin \omega_n \sqrt{1 \quad \zeta^2}t}{\sqrt{1 \quad \zeta^2}} \quad e^{\ \zeta\omega_n t} \quad \frac{N \quad \omega}{K_{VCO}} \cos \omega_n \sqrt{1 \quad \zeta^2}t \quad e^{\ \zeta\omega_n t}$$

$$+ \frac{N \quad \omega}{K_{VCO}} \quad \zeta < 1 \tag{10.52}$$

These expressions are plotted in Figure 10.17. It is interesting to note that, from this expression, it looks like high ζ is best for fast settling; however, it should be noted that even though the frequency appears to lock quickly, there is still a long period before the system is phase locked. Therefore, although these plots may be

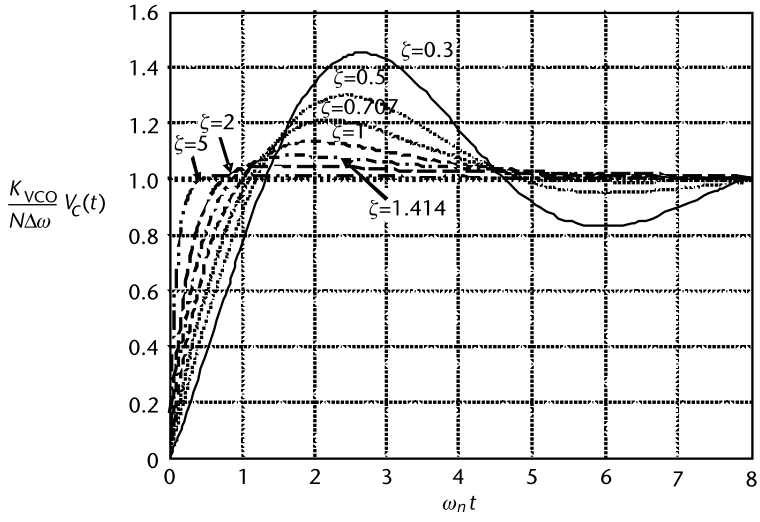

Figure 10.17 Control voltages for frequency step, high-gain second-order loop.

useful to compare with the control voltage (which is more readily available from simulation), they can also be misleading.

Example 10.1: Limits of the Theory So Far

Assume that a synthesizer is designed with a charge pump, PFD, and loop filter like the ones considered thus far in this chapter. Assume that the loop filter is designed so that the system has a damping constant of 0.707 and a 3-dB bandwidth of 150 kHz. What is the maximum frequency step at the input so that the theory so far is still able to predict the behavior of the system? Provided this condition is met, how long will it take the system to settle from a frequency step?

Solution:
This is a classic question that the authors have been using to torment undergraduate students for years now. Besides being good exam fodder, it also illustrates a very important point.

First, we compute the natural frequency of the loop using approximation #2 from (10.28):

$$\omega_n \quad \frac{\omega_{3\text{dB}}}{\left(1+\zeta\sqrt{2}\right)} = \frac{2\pi\times 150\text{ kHz}}{2} = 2\pi\times 75\text{ kHz} \quad \text{(approximation \#2)}$$

Now referring to Figure 10.16, the maximum normalized phase error to a frequency step is about 0.46 for $\zeta = 0.707$. Therefore, the maximum phase error is:

$$\theta_{e\_\max} = 0.46\frac{\omega}{\omega_n}$$

The maximum phase error that the PFD can withstand is 2π. Therefore, the largest frequency step that the system can handle is:

$$\omega_{\max} = \frac{\theta_{e\_\max}\omega_n}{0.46} = \frac{2\pi(2\pi\times 75\text{ kHz})}{0.46} = 6.43\frac{\text{Mrad}}{s} = 1.02\text{ MHz}$$

If the frequency step is any larger than this, then the PLL will lose lock and cycle slip, and the transient response will no longer look like Figure 10.16. If it is less than this, Figure 10.16 should do a fair job of predicting the result. In this case it will take a normalized time of $\omega_n t = 7$ before the transient settles, or about 14.9 μs.

10.6.2 Nonlinear Transient Behavior

When a PLL is first turned on or if it experiences a large frequency step at the input, then it may lose lock. In this case, the linear control theory that has been used so far will not apply, as nonlinearities are involved in lock acquisition. The main source of nonlinearity is the finite linear range of the phase detector. Additionally, there is a finite range over which the loop can acquire lock because VCOs have a finite tuning range. If the loop attempts to lock the VCO to a frequency outside its

range, then the loop will never acquire lock. In addition, if the loop has a finite dc gain, then the range of lock acquisition may also be limited by the finite range of the phase detector.

In general, a frequency step will result in a nonzero phase error. The general transfer function for phase error is:

$$\frac{\theta_e}{\theta_R} = \frac{s}{s + KF(s)} \tag{10.53}$$

Now if a frequency step is applied to this system, the steady-state phase error will be:

$$\theta_{e\_ss} = \lim_{s \to 0} \left[\frac{\omega}{s^2} \div \times \frac{s}{s + KF(s)} \div \times s \right] = \frac{\omega}{KF(0)} \tag{10.54}$$

Now in the second-order charge pump and PFD system, the steady-state phase error will always be zero because there is an integrator in $F(s)$ and $F(0)$ will go to in (10.54). In other loops without an integrator in $F(s)$, the phase error would be finite. When the steady-state error exceeds the linear range of the phase detector, the loop will lose lock.

However, in the case of the charge pump and PDF loop, this is not an issue, since in lock, the steady-state phase error is always zero. In this case, the locking range is determined exclusively by the VCO.

So if the loop is going to experience a transient frequency step, then how long does it take the loop to reacquire lock? In this situation, the loop goes into a frequency acquisition mode and the output of the PFD (in the frequency detection mode) will look something like that shown in Figure 10.7. In this case, the charge pump will put out pulses of current of value I, which vary in width between almost a complete reference period and a width of almost zero. Therefore, the average current produced by the charge pump until the loop acquires lock will be approximately $I/2$. If it is assumed that all of this current flows onto the capacitor C_1, then the change in voltage across the capacitor as a function of time will be:

$$\frac{v_C}{t} = \frac{I}{2C_1} \tag{10.55}$$

Therefore, the settling time will be:

$$T_s = \frac{2 \quad v_C C_1}{I} \tag{10.56}$$

From this equation and making use of the relationship in (10.4) and (10.20), the settling time can be determined for an input frequency change ω as:

$$T_s = \frac{2C_1 \quad \omega N}{IK_{\mathrm{VCO}}} = \frac{\omega}{\pi \omega_n{}^2} \tag{10.57}$$

It should be noted that v_C and K_{VCO} relate to the change in VCO frequency, not to the change of input frequency. This leads to a factor of N in the equation.

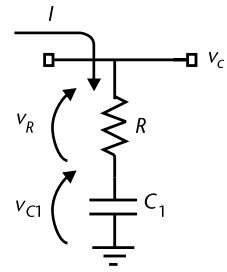

Figure 10.18 Simplified loop filter to illustrate settling behavior.

Thus, one can see directly that as the loop bandwidth is expanded, the settling time will decrease as expected. Even though this is the main result in which we are interested, a few more details of the transient behavior of the control voltage are very interesting to examine. To start this discussion, we will assume that the loop filter does not have a capacitor C_2 and is currently charging up towards lock as shown in Figure 10.18.

In this case, the charge pump current will alternately turn on and turn off. When the charge pump current is off, then v_R will be zero and the control voltage v_c will be equal to the voltage across the capacitor v_{C1}. However, when the charge pump current is on, then v_c will be equal to $v_c = v_{C1} + IR$, where IR is the voltage drop across the resistor. This is illustrated in Figure 10.19.

When the capacitor C_2 is included, its filtering effect results in the behavior being a little more complicated. In this case, when the charge pump is on, most of the current still flows into C_1, which still happily charges towards lock, but when it turns off, there is no longer an instantaneous change in v_c. In this case, C_2 keeps v_c high, and current flows from C_2 back into C_1 through R. This is shown in Figure 10.20.

Thus, the presence of C_2 tends to smooth out the ripple on the control voltage. For context, the same plot as the one in Figure 10.19 is shown in Figure 10.21(b) where voltage waveforms are also plotted over a larger percentage of acquisition in Figure 10.21(a).

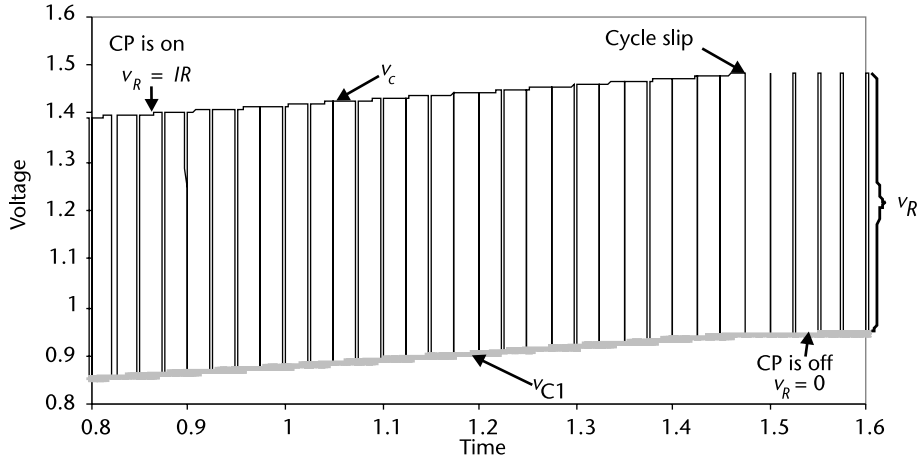

Figure 10.19 Example showing the voltages on the loop filter during acquisition (no C_2 present).

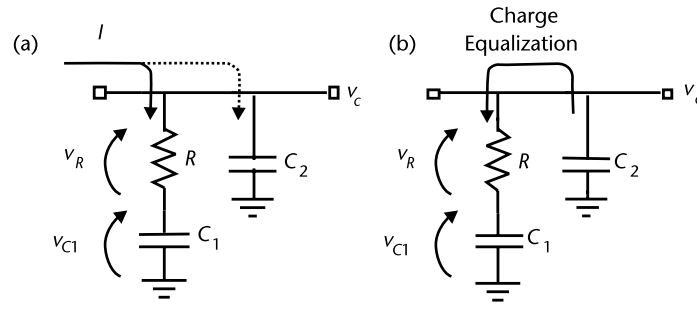

Figure 10.20 Illustration of loop filter behavior (C_2 present). (a) Charge pump is on and charging both C_1 and C_2. (b) Charge pump is off and C_2 is discharging into C_1.

Example 10.2: Simulation and Estimation of Loop Settling Times

A 3.7–4.3-GHz synthesizer with a step size of 1 MHz is required. A 40-MHz crystal oscillator, a charge pump with a $2\pi \times 100$ μA output current, and a VCO (operating from a 3-V supply) are available. Design a fractional N synthesizer with a loop bandwidth of 150 kHz using these components. Estimate the settling time of the loop for a 30-MHz and a 300-MHz frequency step. Simulate and compare.

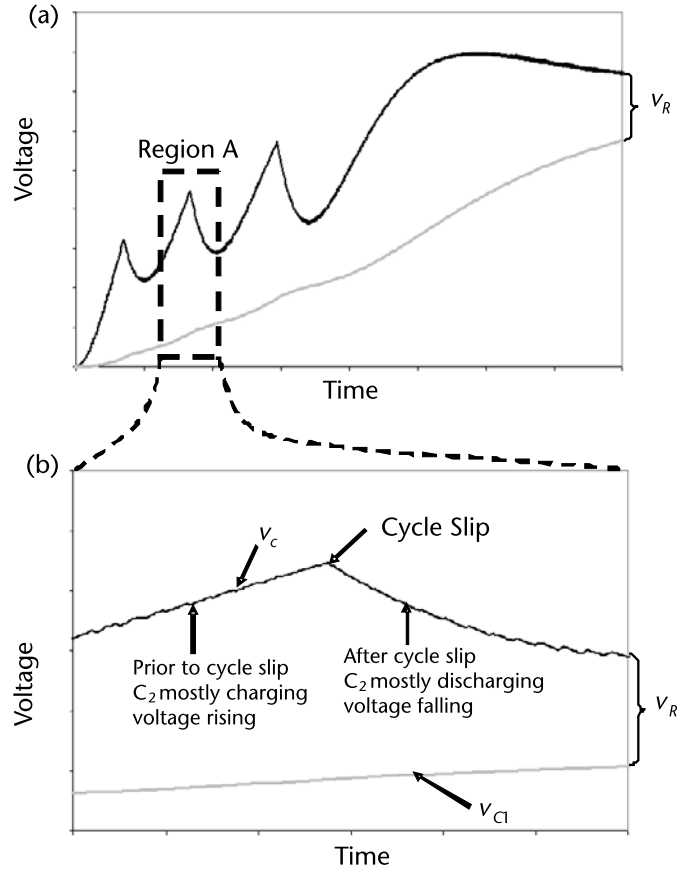

Figure 10.21 Example showing the voltages on the loop filter during acquisition (C_2 present): (a) complete settling and (b) zoom in on region A.

Solution:
First, if the VCO is operating with a 3-V supply and must have a 600-MHz tuning range, we can estimate that its K_{VCO} will be 200 MHz/V. In addition, since we know the charge pump current, we know that that the K_{phase} will be 100 μA/rad. For a VCO with a nominal frequency of 4 GHz and a reference frequency of 40 MHz, the division ratio will be 100. The next step is to size the loop filter. A 3-dB frequency of 150 kHz requires a natural frequency of 75 kHz (see Example 10.1). Thus, components can be determined as:

$$C_1 = \frac{IK_{VCO}}{2\pi \times N\omega_n^2} = \frac{2\pi \times 100\ \mu\text{A} \times 2\pi \times 200 \dfrac{\text{MHz}}{V} \div}{2\pi \times 100(2\pi \times 75\ \text{kHz})^2} = 5.66\ \text{nF}$$

In order to set R, we need to pick a damping constant. Let us pick $1/\sqrt{2}$, or 0.707, which is a popular choice. Now:

$$R = 2\zeta\sqrt{\frac{2\pi \times N}{IK_{VCO}C_1}} = 2\ \frac{1}{\sqrt{2}} \div \sqrt{\frac{2\pi \times 100}{2\pi \times 100\ \mu\text{A}\ \ 2\pi \times 200 \dfrac{\text{MHz}}{\text{V}} \div \times 5.66\ \text{nF}}} = 530$$

and we will set $C_2 = 566$ pF at one-tenth the value of C_1.

Now, a step in output frequency of 30 MHz and 300 MHz corresponds to a step in the reference frequency of 0.3 MHz and 3 MHz, respectively. We learned in Example 10.1 that the maximum input frequency step that can be tolerated for a system with these parameters is 1 MHz. Therefore, the first frequency step will be a linear one and the output will follow the theory of the previous section. We expect that it will take approximately 15 μs to settle, as we discovered previously.

In contrast, the second frequency step will involve cycle slipping. For this nonlinear case, we use the formula given in (10.57) to estimate the acquisition time as:

$$T_s = \frac{\omega}{\pi\omega_n^2} = \frac{2\pi \times 3\ \text{MHz}}{\pi(2\pi \times 75\ \text{kHz})^2} = 27\ \mu\text{s}$$

Therefore, complete settling in this case should take 27 μs plus an additional 15 μs for phase lock.

This behavior can be simulated using ideal components in a simulator such as Cadence's Spectre. The blocks for the divider, VCO, PFD, and charge pump can be programmed using ideal behavioral models. These can be connected to the loop filter that has just been designed. From these simulations, we can look at the control voltage on the VCO to verify the performance of the loop. A plot of the response of the system to a 0.3-MHz step at its input, compared to the simple theory is shown in Figure 10.22. From this graph, it is easy to see that the simple theory does an excellent job of predicting the settling behavior of the loop with only a slight deviation. This small discrepancy is most likely due to the presence of C_2 and to the sampling nature of the loop components.

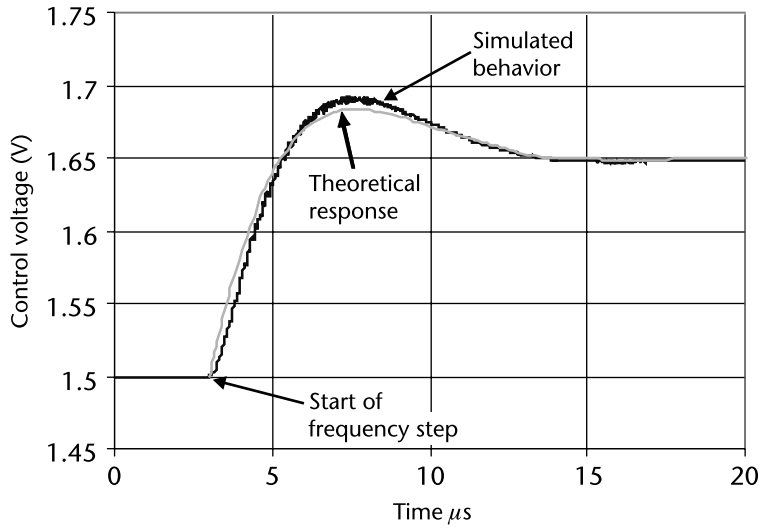

Figure 10.22 Response of the PLL design's control voltage during a 30-MHz frequency step.

The second frequency step can be simulated as well. The results of this simulation are plotted in Figure 10.23 and compared to the linear voltage ramp suggested by simple theory previously. In this case, this plot shows that the nonlinear response is predicted fairly well by the simple formula; however, the actual response is slightly faster. The tail of this graph is cut off, but the loop settled in about 39 μs, which is very close to the 42 μs predicted. The main difference between the simple estimate and reality is the fact that phase acquisition begins before the PLL actually reaches its final frequency value. We predicted it would begin in this simulation at 30 μs (when the simple theory predicts that the voltage ramp will reach its final value), but the linear portion of the graph actually starts earlier than this, at about 25–26 μs. This accounts for our slightly pessimistic estimate. Still, such a simple estimate is remarkably good at predicting quite complicated behavior.

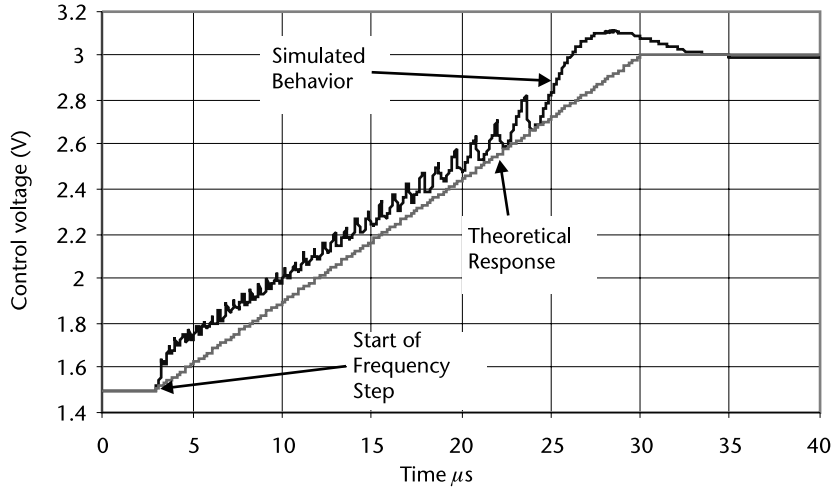

Figure 10.23 Response of the PLL design's control voltage during a 300-MHz frequency step.

10.6.3 Various Noise Sources in PLL Synthesizers

10.6.3.1 VCO Noise

All the circuits in the synthesizer contribute to the overall noise in different ways and the noise they produce has different characteristics. For instance, the phase noise from a VCO can be described as [60]:

$$\varphi^2_{\mathrm{VCO}}(\omega) = \frac{C}{\omega^2} + D \tag{10.58}$$

where C is a constant of proportionality and ω is the frequency offset from the carrier. Thus, at most frequencies of interest, the phase noise produced by the VCO will go down at 20 dB/decade as we move away from the carrier. This will not continue indefinitely, as thermal noise will put a lower limit on this phase noise D which, for most integrated VCOs, is somewhere between –120 and –150 dBc/Hz. VCO phase noise is usually dominant outside the loop bandwidth and of less importance at low offset frequencies.

10.6.3.2 Crystal Reference Noise

Crystal resonators are widely used in frequency control applications because of their unequaled combination of high Q, stability and small size. The resonators are classified according to "cut," which is the orientation of the crystal wafer (usually made from quartz) with respect to the crystallographic axes of the material. The total noise power spectral density of a crystal oscillator can be found by Leeson's formula [64]:

$$\varphi^2_{XTAL}(\omega) = 10^{16\pm1} \times 1 + \left(\frac{\omega_0}{2\omega \times Q_L}\right)^2 \left(1 + \frac{\omega_c}{\omega}\right) \tag{10.59}$$

where ω_0 is the oscillator output frequency, ω_c is the corner frequency between $1/f$ and thermal noise regions, which is normally in the range 1~10 kHz, and Q_L is the loaded quality factor of the resonator. Since Q_L for crystal resonator is very large (normally in the order of 10^4 to 10^6), the reference noise contributes only to the very close-in noise and it quickly reaches thermal noise floor at an offset frequency around ω_c.

10.6.3.3 Frequency Divider Noise

Frequency dividers consist of switching logic circuits, which are sensitive to clock timing jitter. The jitter in the time domain can be converted to phase noise in the frequency domain. Time jitter or phase noise occurs when rising and falling edges of digital dividers are superimposed with spurious signals such as thermal noise and flicker noise in semiconductor materials. Ambient effects result in variation of the triggering level due to temperature and humidity. Frequency dividers generate spurious noise especially for high frequency operation. While there is no substitute for real simulations or measurements of a particular divider, Kroupa provided an empirical formula that describes the amount of phase noise that typical frequency

dividers add to a signal [16, 17]:

$$\varphi^2_{\text{Div\_Added}}(\omega) \quad \frac{10^{\ 14\pm1} + 10^{\ 27\pm1}\omega_{do}^2}{2\pi\times\ \omega} + 10^{\ 16\pm1} + \frac{10^{\ 22\pm1}\omega_{do}}{2\pi} \tag{10.60}$$

where ω_{do} is the divider output frequency and ω is the offset frequency. Notice that the first term in (10.60) represents the flicker noise and the second term gives the white thermal noise floor. The third term is caused by timing jitter due to coupling, ambient, and supply variations. Additional information may be found in [2].

10.6.3.4 Phase Detector Noise

Phase detectors experience both flicker and thermal noise. Although the noise generated by phase detectors depends on a number of factors such as circuit topology and the technology used to implement them, a rule of thumb is that at large offsets, phase detectors generate a white phase noise floor of about 160 dBc/Hz, which is thermal noise dominant. The noise power spectral density of phase detectors is given by [55]:

$$\varphi^2_{\text{PD}}(\omega) \quad \frac{2\pi\times10^{\ 14\pm1}}{\omega} + 10^{\ 16\pm1} \tag{10.61}$$

Additional information may be found in [2].

10.6.3.5 Charge Pump Noise

The noise of the charge pump can be characterized as an output noise current and is usually given in $\text{pA}/\sqrt{\text{Hz}}$. Note that at this point in the loop the phase is represented by the current. The charge pump output current noise can be a strong function of the frequency and of the width of the current pulses; therefore, for low noise operation it is desirable to keep the charge pump currents matched as well as possible. This is because current sources only produce noise when they are on. In an ideal loop, when locked, the charge pump is always off. However, nonidealities result in finite pulses, but the closer reality matches the ideal case, the less noise will be produced. Also note that as the frequency is decreased, the 1/*f* noise will become more important, causing the noise to increase. This noise can often be the dominant noise source at low frequency offsets. Charge pump noise can be simulated with proper tools and the results depend on the design in question so no simple formula will be given here.

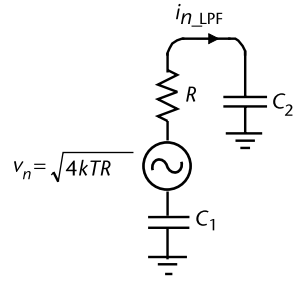

Figure 10.24 Loop filter with thermal noise added.

10.6.3.6 Loop Filter Noise

Loop filters are simple RC circuits and can be analyzed for noise in the frequency domain in a linear manner. The most common loop filter that has been examined in this chapter will now be analyzed. It contains only one noise source, the thermal noise associated with R. Thus, the loop filter with associated noise can be drawn as shown in Figure 10.24. Now the noise voltage develops a current flowing through the series combination of C_1, C_2, and R (assuming that the charge pump and VCO are both open circuits), which is given by:

$$i_{n\_\mathrm{LPF}} = \frac{1}{R} \times \frac{v_n s}{s + \dfrac{C_1 + C_2}{C_1 C_2 R}} \quad \frac{1}{R} \times \frac{v_n s}{s + \dfrac{1}{C_2 R}} \tag{10.62}$$

Thus, this noise current will have a highpass characteristic, and therefore the loop will not produce any noise at dc and this noise will increase with frequency until the highpass corner is reached, after which it will be flat.

10.6.3.7 Noise

If a noise shaper is used to control the divide ratio in a synthesizer, then it will also generate phase noise. Fortunately, discrete fractional spurs become more like random noise after noise shaping, which is the main reason these more complicated circuits are preferred over the simple accumulators discussed later in this chapter. The single-sideband phase noise of the modulator is given by:

$$\begin{aligned} \frac{\varphi^2 \quad (f)[\mathrm{rad}^2/\mathrm{Hz}]}{2} &= \frac{(2\pi)^2}{24 f_r} \times 2\sin \quad \frac{\pi f}{f_r} \div \quad ^{2(m \ 1)} \\ PN \quad (f)[\mathrm{dBc}/\mathrm{Hz}] &= 10\log \quad \frac{(2\pi)^2}{24 f_r} \quad 2\sin \quad \frac{\pi f}{f_r} \div \quad ^{2(m \ 1)} \end{aligned} \tag{10.63}$$

where f is the offset frequency, f_r is the reference sampling frequency, and m is the order of the modulator. Note that the noise is shaped by the modulator. The higher the order of the modulator, the more noise is pushed to higher frequencies where it can be filtered by the PLL.

10.6.4 In-Band and Out-of-Band Phase Noise in PLL Synthesis

The noise transfer functions for the various noise sources in the loop can be derived quite easily using the theory already presented. There are three noise-transfer functions: one for the VCO, one for all other sources of noise in the loop, and one for the modulator. All noise generated by the PFD, charge pump, divider, and loop filter is referred back to the input, and the noise from the VCO is referred to the output as shown in Figure 10.25. The transfer function for $\varphi_{\mathrm{noiseI}}(s)$ has already been derived (as it is the same as the loop transfer function in the continuous domain) and is given by:

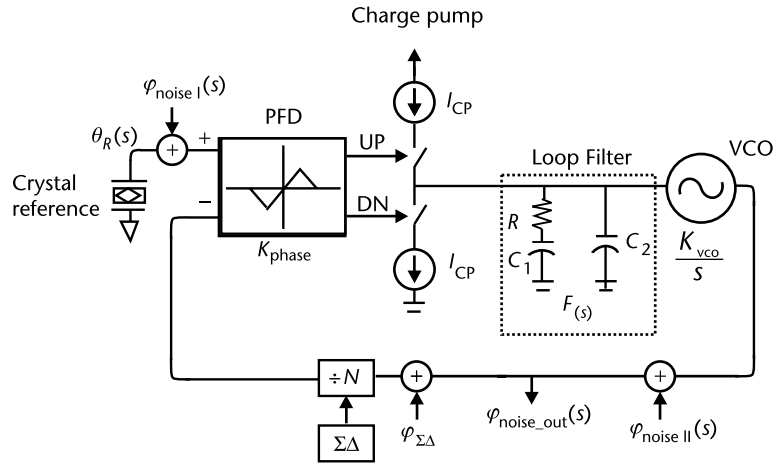

Figure 10.25 A synthesizer showing places where noise is injected.

$$\frac{\varphi_{\text{noise out}}(s)}{\varphi_{\text{noiseI}}(s)} = \frac{F(s)K_{\text{VCO}}K_{\text{phase}}}{s + \dfrac{F(s)K_{\text{VCO}}K_{\text{phase}}}{N}} \tag{10.64}$$

where for the charge pump PFD loop, the transfer function for the filter, divider, and crystal reference noise becomes:

$$\frac{\varphi_{\text{noise out}}(s)}{\varphi_{\text{noiseI}}(s)} = \frac{\dfrac{IK_{\text{VCO}}}{2\pi \times C_1}(1 + RC_1 s)}{s^2 + \dfrac{IK_{\text{VCO}}}{2\pi \times N} R_S + \dfrac{IK_{\text{VCO}}}{2\pi \times NC_1}} \tag{10.65}$$

This function is a lowpass filter. This means that for low frequencies (inside the loop bandwidth), the loop will track the input phase (which includes the phase noise), and thus this noise will be transferred to the output. At higher offset frequencies (outside the loop bandwidth), this noise is suppressed as the loop prevents the VCO from following these changes in phase. Also note that the division ratio plays a very important part in this transfer function. It can be seen that at low reference frequencies, where the s and s^2 terms in (10.65) can be ignored relative to the constant terms, the higher division ratio N directly results in higher phase noise. This is one of the strongest arguments for using fractional-N architectures in synthesizers, as with large division ratios, it is hard to get a low phase noise performance.

The transfer function for the noise due to the VCO is slightly different. In this case, the input noise is set to zero and then the transfer function is derived in the usual way. It is given in general by:

$$\frac{\varphi_{\text{noise out}}(s)}{\varphi_{\text{noise II}}(s)} = \frac{s}{s + \dfrac{F(s)K_{\text{VCO}}K_{\text{phase}}}{N}} \tag{10.66}$$

Using our loop will give:

$$\frac{\varphi_{\text{noise out}}(s)}{\varphi_{\text{noise II}}(s)} = \frac{s^2}{s^2 + \frac{IK_{\text{VCO}}}{2\pi \times N} RS + \frac{IK_{\text{VCO}}}{2\pi \times NC_1}} \tag{10.67}$$

This is a highpass filter. Thus, at low offsets inside the loop bandwidth, the VCO noise is suppressed by the feedback action of the loop, but outside the loop bandwidth, the VCO is essentially free running and the loop noise approaches the VCO noise.

This noise from the is injected into the system in a third location as shown in Figure 10.25. Therefore, its noise transfer function to the output is given by:

$$\frac{\varphi_{\text{noise\_out}}(s)}{\varphi \quad (s)} = \frac{\frac{F(s)K_{\text{VCO}}K_{\text{phase}}}{N}}{s + \frac{F(s)K_{\text{VCO}}K_{\text{phase}}}{N}} \tag{10.68}$$

Note that, due to the highpass nature of the noise transfer function, the order of the loop rolloff is very important. It can be seen from (10.63) that the noise of an mth-order modulator has a noise shaping slope of $20(m - 1)$ dB/decade, while an nth-order lowpass loop filter has a rolloff slope of $20n$ dB/decade. Therefore, the order of the loop filter must be higher than or equal to the order of the modulator in order to attenuate the out-of-band noise due to the modulation. Thus, for instance, when calculating the effect of the modulator on out-of-band noise, it is necessary to include the effect of C_2 in the loop filter, as this will provide extra attenuation out of band. In this case the noise transfer function to the output would be:

$$\frac{\varphi_{\text{noise\_out}}(s)}{\varphi \quad (s)} = \frac{K_{\text{VCO}}K_{\text{phase}}(1 + sC_1R)}{s^2N(C_1 + C_2)(1 + sC_sR) + K_{\text{VCO}}K_{\text{phase}}(1 + sC_1R)} \tag{10.69}$$

where $C_s = \dfrac{C_1C_2}{C_1 + C_2}$.

Example 10.3: System Phase Noise Calculations

Estimate the phase noise for the synthesizer that was designed in Example 10.2. The VCO has a phase noise of –120 dBc/Hz at a 1-MHz offset (it bottoms out at –130 dBc/Hz), and the charge pump puts out a noise current of 10 $\text{pA}/\sqrt{\text{Hz}}$. Ignoring PFD, the divider, modulator, and reference noise sources plot the phase noise. In addition, what would be the phase noise of an equivalent integer-N design?

Solution:

From Example 10.2, we know the charge pump current and we know that that the K_{phase} will be 100 μA/rad. Now in the case of the integer-N synthesizer the reference

must be 1 MHz in order to get a step size of 1 MHz. Therefore, the division ratio will be 4,000. Knowing that we want a loop bandwidth of 150 kHz means that we need a natural frequency of 75 kHz (assuming a damping constant of 0.707), and this means that for the integer-N design, C_1 and R are 141.5 pF and 21.2 k , respectively. Now we will assume that the VCO follows the 20 dB/decade rule just outlined. Therefore, we can come up with an expression for the phase noise of the VCO based on (10.58):

$$C = \log^{-1}\left(\frac{PN_{\text{VCO}}}{10}\right) \times \omega^2 = \log^{-1}\left(\frac{120}{10}\right) \times (2\pi \times 1\text{ MHz})^2 = 39.5\frac{\text{rad}^4}{\text{Hz}^2}$$

Now, since the VCO bottoms out at –130 dBc/Hz, this means that we can determine the D term of the VCO phase noise equation (10.58):

$$D = \log^{-1}\left(\frac{PN_{\text{VCO}}}{10}\right) = \log^{-1}\left(\frac{130}{10}\right) = 10^{13}\frac{\text{rad}^2}{\text{Hz}}$$

This can be turned into an equation that has units of voltage instead of units of V^2:

$$\varphi_{\text{VCO}}(\omega) = \sqrt{\frac{39.5}{\omega^2} + 10^{13}}\frac{\text{rad}}{\sqrt{\text{Hz}}}$$

The output noise current from the charge pump can be input referred by dividing by K_{phase}:

$$\text{Noise}_{\text{CP}} = \frac{i_n}{K_{\text{phase}}} = \frac{10\dfrac{p\text{A}}{\sqrt{\text{Hz}}}}{100\dfrac{\mu\text{A}}{\text{rad}}} = 100n \times \frac{\text{rad}}{\sqrt{\text{Hz}}}$$

The noise from the loop filter must also be moved back to the input:

$$\text{Noise}_{\text{LPF}}(\omega) = \frac{1}{K_{\text{phase}}}\left|\frac{\sqrt{\dfrac{4kT}{R}}j\omega}{j\omega + \dfrac{1}{C_2R}}\right|$$

Clearly, noise from the lowpass filter is dependent on filter component values as well as the phase detector gain. Now input-referred noise from the loop filter and the charge pump can each be substituted into (10.65) while noise from the VCO can be substituted into (10.67) to determine the contribution to the phase noise at the output. Once each component value at the output is calculated, the overall noise can be computed (noting that noise adds as power). So, for instance, in the case of the noise due to the charge pump, the output phase noise for the fractional-N case is [from (10.65)]:

$$\varphi_{\text{noise out\_CP}}(s) = \frac{2.22 \times 10^{13}(1 + 3 \times 10^{-6}s)}{s^2 + 6.66 \times 10^5 s + 2.22 \times 10^{11}} 100n \times \frac{\text{rad}}{\sqrt{\text{Hz}}}$$

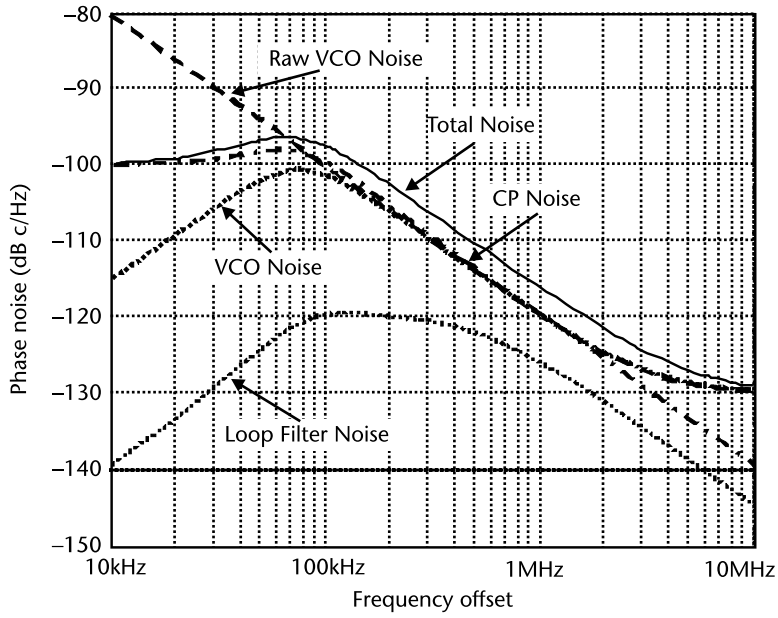

Figure 10.26 Phase noise of various components and overall phase noise for the system with fractional-*N* divider.

Therefore, to plot phase noise in dBc/Hz, we take:

$$PN_{\mathrm{CP}}(\omega) = 20\log\left|\frac{2.22\times10^{13}(1+3\times10^{-6}j\omega)}{(j\omega)^2+6.66\times10^{5}j\omega+2.22\times10^{11}}\right|100n\times\frac{\mathrm{rad}}{\sqrt{\mathrm{Hz}}}\div$$

The results of this calculation and similar ones for the other noise sources are shown in Figure 10.26. The total phase noise is computed by:

$$\varphi_{\mathrm{total}} = \sqrt{\varphi_{\mathrm{noise\_out\_CP}}{}^2 + \varphi_{\mathrm{noise\_out\_VCO}}{}^2 + \varphi_{\mathrm{noise\_out\_LPF}}{}^2}$$

and is shown in the figure.

If we assume that the numbers given so far have been for single sideband phase noise, then we can also compute the integrated phase noise of this design as:

$$\mathrm{IntPN}_{\mathrm{rms}} = \frac{180\sqrt{2}}{\pi}\sqrt{\int_{f=10\ \mathrm{kHz}}^{f=10\ \mathrm{MHz}} \varphi_{\mathrm{total}}{}^2 df} = 0.41^\circ$$

Note, in this example, that the loop-filter noise is quite low and could have been ignored safely. Also note that due to the frequency response of the filter even in-band, noise from the loop filter is attenuated at lower frequencies. Inside of the loop bandwidth, the total noise is dominated by the charge pump, which is the more dominant in-band noise source. Note that out of band, the noise is slightly higher than the VCO noise. This is because the charge pump is still contributing. This could be corrected by making the loop bandwidth slightly smaller and thus attenuating the out-of-band contribution of the charge pump by a few more decibels.

With the integer-*N* numbers, the phase noise is shown in Figure 10.27. Note that with integer *N*, the noise is completely dominated by the charge pump, both

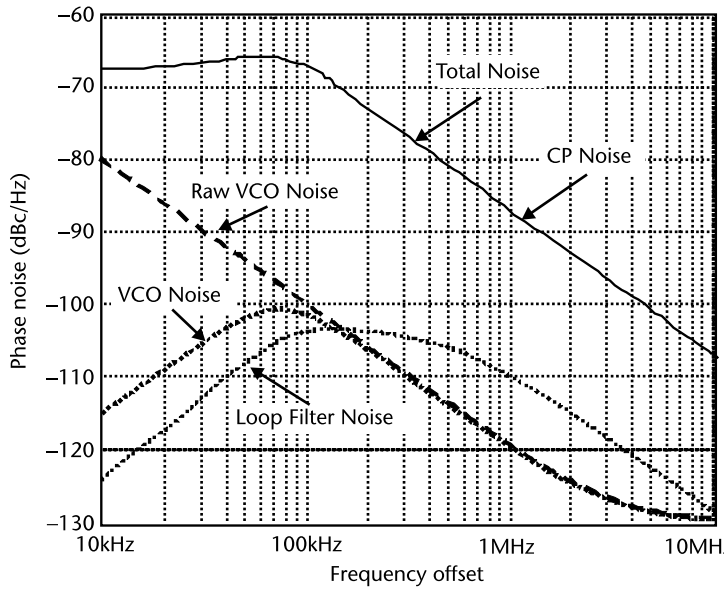

Figure 10.27 Phase noise of various components and overall phase noise for the system with integer-*N* divider.

inside and outside of the loop bandwidth. In order to reduce the effect of charge pump noise, the loop bandwidth in this case should be reduced by at least two orders of magnitude. Note also the dramatic change in the in-band phase-noise performance between the two designs. While the fractional design has –100 dBc/Hz of in-band noise, this design has a performance of only –67 dBc/Hz. Note that the two numbers are different by 20log(40), which is the ratio of the two divider values, as would be expected.

10.7 Fractional-*N* PLL Frequency Synthesizers

In contrast to an integer-*N* synthesizer, a fractional-*N* synthesizer allows the PLL to operate with high reference frequency while achieving a fine step size by constantly swapping the loop division ratio between integer numbers. As a result, the average division ratio is a fractional number. As has been shown, a higher reference frequency leads to a lower in-band phase noise and a faster PLL transient response. In addition, for multiband applications, often the channel spacing of the different bands is skewed, requiring an even lower reference frequency if an integer *N* synthesizer is to cover both bands.

Example 10.4: The Problem with Using an Integer-*N* Synthesizer for Multiband Applications

Determine the maximum reference frequency of an integer-*N* frequency synthesizer required to cover channels from 2,400 MHz to 2,499 MHz spaced 3 MHz apart and channels from 5,100 MHz to 5,200 MHz spaced 4 MHz apart.

Solution:
If an integer-*N* synthesizer were designed to service only one of these bands, then it would have a maximum reference frequency of 3 MHz in the first case and 4 MHz in the second case. However, if a synthesizer must be designed to cover both of these bands, then its step size must be 1 MHz to allow it to tune exactly to every frequency required.

In the simplest case, the fractional-*N* synthesizer generates a dynamic control signal that controls the divider, changing it between two integer numbers. By toggling between the two integer division ratios, a fractional division ratio can be achieved by time-averaging the divider output. As an example, if the control changes the division ratio between 8 and 9, and the divider divides by 8 for 9 divider output cycles and by 9 for 1 divider output cycle and then the process repeats itself, then the average division ratio will be:

$$\overline{N} = \frac{\text{Total Divider Input Pulses Needed}}{\text{Total Divider Output Pulses Generated}} = \frac{8 \quad 9+9 \quad 1}{10} = 8.1 \tag{10.70}$$

If the divider were set only to divide by 8, then it would produce 10 output pulses for 80 input pulses. However, now it will take 81 input pulses to produce 10 output pulses. Thus, the device swallows one extra input pulse to produce every 10 output pulses. In the PLL synthesizer, this time average is dealt with by the transfer function in the loop. This transfer function will always have a lowpass characteristic. Thus, it will deliver the average error signal to the VCO. As a result, the output frequency will be the reference frequency multiplied by the average division ratio. However, toggling the divider ratio between two values in a repeating manner generates a repeating time sequence. In the frequency domain, this periodic sequence

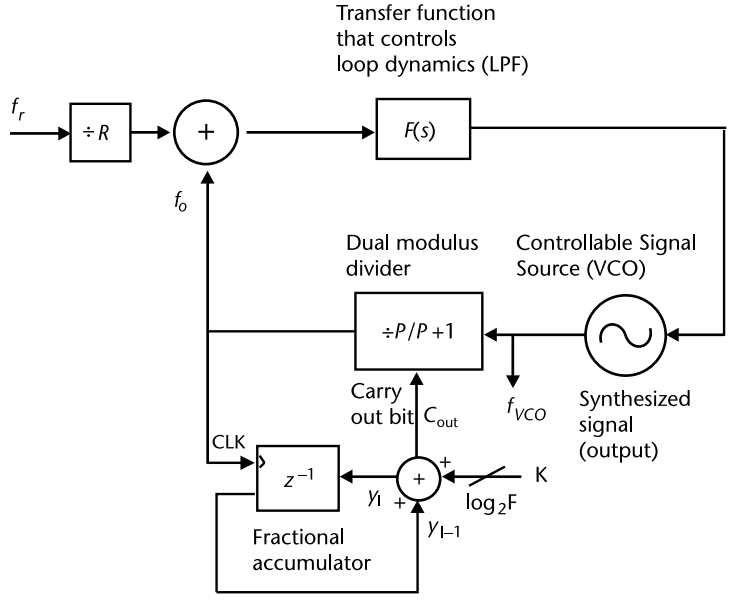

Figure 10.28 A fractional-*N* frequency synthesizer with a dual modulus prescaler.

will generate spurious components (or spurs) at integer multiples of the repetition rate of the time sequence. Such spurious components can be reduced by using modulators in which the instantaneous division ratio is randomized while maintaining the correct average value.

10.7.1 Fractional-*N* Synthesizer with a Dual Modulus Prescaler

Figure 10.28 illustrates one way to implement a simple fractional-*N* frequency synthesizer with a dual modulus prescaler $P/P+1$. Note that it is called a dual modulus prescaler because it can be programmed to two division ratios. As discussed in the previous section, the fractionality can be achieved by toggling the divisor value between two values, P and $P+1$. The modulus control signal (C_{out}) is generated using an accumulator (also called an integrator or adder with feedback, or a counter) with size of F (or $\log_2 F$ bits). That is, an overflow occurs whenever the adder output becomes equal to or larger than F. At the ith clock rising edge, the accumulator's output y_i can be mathematically expressed as:

$$y_i = (y_{i\ 1} + K_i) \bmod F \tag{10.71}$$

where $y_{i\ 1}$ is the output on the previous rising clock edge, and K_i is a user-defined input and its value will determine the fractional divider value. Its use will be illustrated shortly in Example 10.5. Note that the modular operation (A mod B) returns the remainder of (A B) and is needed for modeling the accumulator overflow.

Example 10.5: A Simple Accumulator Simulation

Describe the operation of a 3-bit accumulator with input $K = 1$ and $K = 3$, assuming the accumulator seed value (i.e., the initial accumulator output value) is equal to 0.

Solution:
If the accumulator has 3 bits, the size of the accumulator is $2^3 = 8$, or $F = 8$, even though the largest value which can be stored is 111, corresponding to 7. If input word $K = 1$, namely, 001, the accumulator value y increases by 1 every cycle until it reaches the maximum value that can be represented using 3 bits, namely, $y_{\max} = 7 = 111$. After this point, the accumulator will overflow, leaving its value $y = 0$ and $C_{\text{out}} = 1$. It will take 8 clock cycles for the accumulator to overflow if $K = 1$. In other words, the accumulator size $F = 8$. For $K = 3 = 011$, the accumulator adds an increment value of 3 every cycle and thus overflows more often. The accumulator value and its carryout C_{out} are summarized cycle by cycle in Table 10.1 for $K = 1$ and Table 10.2 for $K = 3$.

Table 10.1 Accumulator Operations with $F = 8$, $K = 1$

Clock cycle i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
y_i	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2
$y_{i\ 1}$	NA	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1
C_{out}	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0

Table 10.2 Accumulator Operations with $F = 8$, $K = 3$

Clock cycle i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
y_i	0	3	6	1	4	7	2	5	0	3	6	1	4	7	2	5	0	3	6
$y_{i\ 1}$	NA	0	3	6	1	4	7	2	5	0	3	6	1	4	7	2	5	0	3
C_{out}	1	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0

As shown in the above example, for the $K = 1$ case, C_{out} is high for one cycle and low for seven cycles within every eight clock cycles, so the frequency of C_{out} is $f_{clk}/8$. For the $K = 3$ case, C_{out} is high for three cycles and low for five cycles within every eight clock cycles, so the frequency of C_{out} is $3f_{clk}/8$. In general, for a constant input word K, the accumulator carryout will be high for K cycles and will be low for F-K cycles. Also note that the frequency of C_{out} will be equal to:

$$f_{C_{\text{out}}} = \frac{Kf_{clk}}{F} \tag{10.72}$$

If the dual modulus prescaler divides by P when C_{out} is low and divides by $P + 1$ when C_{out} is high, the average VCO output frequency is:

$$f_{\text{VCO}} = \frac{f_r}{R}\ \frac{(P+1)K + P(F \quad K)}{F}\ = \frac{f_r}{R}\ P + \frac{K}{F} \tag{10.73}$$

Because fractionality is achieved by using this accumulator, it is often called a fractional accumulator. It has a fixed size F due to a fixed number of accumulator bits built into the hardware. The dual modulus prescaler ratio P is normally fixed as well. The only programmable parameter for the architecture shown in Figure 10.28 is the accumulator input K, which can be programmed from 1 to a maximum number of F. Thus, since K is an integer, from (10.73) it can be seen that the step size of this architecture is given by:

$$\text{Step Size} = \frac{f_r}{RF} \tag{10.74}$$

where R is normally fixed to avoid changing the comparison frequency at the input. Note that R is normally as small as possible to minimize the in-band phase noise contribution from the crystal. Thus, step size is inversely proportional to the number of bits ($\log_2 F$), so as a result, the accumulator is normally used to reduce synthesizer step size without increasing R and degrading the in-band phase noise. More detail on spur reduction is given in [2].

10.7.2 Fractional-*N* Synthesizer with Multimodulus Divider

Replacing the dual modulus divider with a multimodulus divider (MMD), the synthesizer architectures shown in Figure 10.28 can be modified to a more generic form, as illustrated in Figure 10.29. Using a multimodulus divider has the advan-

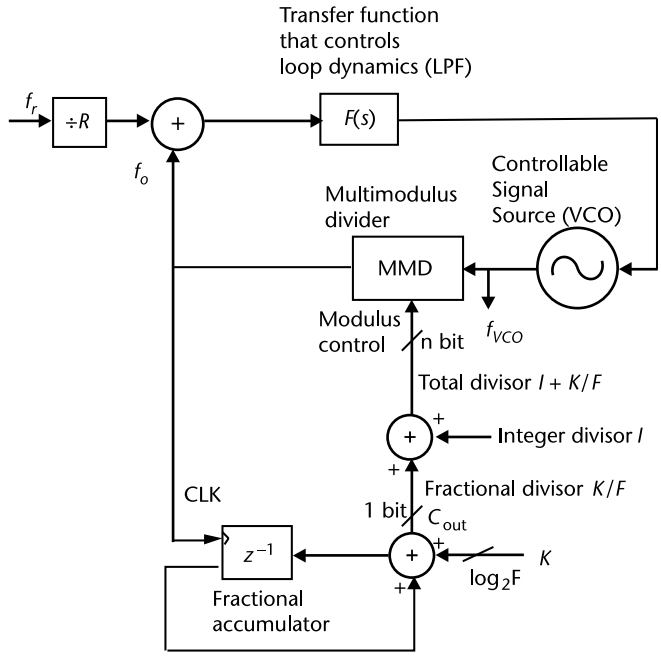

Figure 10.29 A fractional-*N* frequency synthesizer with a multimodulus divider.

tage that the range of frequencies over which the synthesizer can be tuned is expanded, compared to the previous architecture. The synthesizer output frequency is given by:

$$f_{\text{VCO}} = \frac{f_r}{R}\left(I + \frac{K}{F}\right) \tag{10.75}$$

where I is the integer portion of the loop divisor and, depending on the complexity of the design, I could have many possible integer values. For instance, if a loop division ratio 100.25 is needed, we can program $I = 100$, $K = 1$ and $F = 4$. The MMD division ratio is toggled between 100 and 101.

In a popular MMD topology using cascaded 2/3 cells [53] with an n-bit modulus control signal, the MMD division ratio is given by:

$$N_{\text{MMD}} = P_1 + 2^1 P_2 + \ldots 2^{n-2} P_{n-1} + 2^{n-1} P_n + 2^n \tag{10.76}$$

where the MMD programming range is 2^n to $2^{n+1} - 1$. For instance, a 6-bit MMD can be programmed from 64 to 127. The MMD program range can be further extended through the use of additional logic. A wide programming range is critical for multiband frequency synthesis, especially when sigma-delta noise shaping is employed. More details of divider design are given in [2].

10.7.3 Fractional-*N* Spurious Components

The above-discussed fractional-*N* architectures suffer from a common side effect of generating spurious components associated with periodically toggling the loop division ratio. Recall that any repeatable pattern in the time domain causes spurious tones in the frequency domain. The fractional accumulator periodically generates the carryout that toggles the loop division ratio. It is expected that there should be spurious tones at multiples of the carryout frequency $(f_r/R)\times(K/F)$. In the following example, fractional-*N* spurs are analyzed with simulation and measurement.

Example 10.6: The Use of Accumulators in Fractional-*N* Synthesizers

Design a fractional-*N* synthesizer architecture for synthesizing 11 channels from 819.2 MHz to 820.96 MHz with a step size of 160 kHz and reference comparison frequency of $f_r/R = 5.12$ MHz. Determine the frequencies of fractional-*N* spurious components.

Solution:
The synthesizer step size is given by $\frac{f_r}{R}\times\frac{1}{F} = 160$ kHz. Since the comparison frequency is $f_r/R = 5.12$ MHz, the fractional accumulator size can be chosen as:

$$F = \frac{f_r}{R}\times\frac{1}{160\text{ kHz}} = \frac{5{,}120\text{ kHz}}{160\text{ kHz}} = 32$$

which can be implemented using a 5-bit accumulator. The accumulator input (i.e., the fine tune frequency word *K*) can be programmed from 0 to 10 to cover the 11 channels from 819.2 MHz to 820.96 MHz with step size of 160 kHz (the first channel does not require any fractionality). The integer divisor ratio (i.e., the coarse tune frequency word *I*) can be determined by the channel frequency. For instance, the first channel frequency is synthesized as:

$$\frac{f_r}{R}\left(I + \frac{0}{F}\right) = \frac{f_r}{R}\times I = 819.2\text{ MHz}$$

which leads to $I = 160$. Hence, the loop total divisor is given by $N = 160 + K/32$, where $K = 0, 1, \ldots, 10$. If a dual-modulus divider is used to construct a fractional-*N* synthesizer, as illustrated in Figure 10.28, the dual-modulus divider ratio should be $P/P + 1 = 160/161$, which is not the best solution as far as the power consumption and circuit speed is concerned. There are better circuit implementations, such as using a multimodulus divider or a pulse-swallow divider, which allows much of the implemented circuitry to operate at much lower speeds. If a fractional-*N* architecture with MMD is used, as illustrated in Figure 10.29, a 7-bit MMD with a programmable range from 128 to 255 is needed based on (10.76). In any of the above solutions, the loop divisor of the fractional-*N* architecture is toggled between 160 and 161. The modulus control is generated by the accumulator carryout, which has a frequency of $(f_r/R)\times(K/F)$. Thus, the loop is divided by 160 for *K* reference

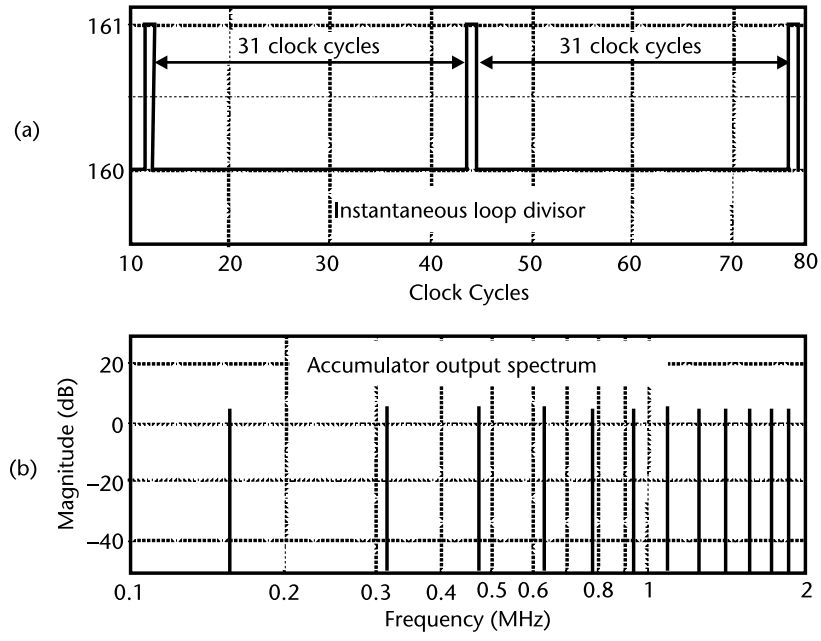

Figure 10.30 (a) Simulated fractional accumulator output with loop divisor $N = 160 + 1/32$ and (b) the comparison frequency $f_r/R = 5.12$ MHz.

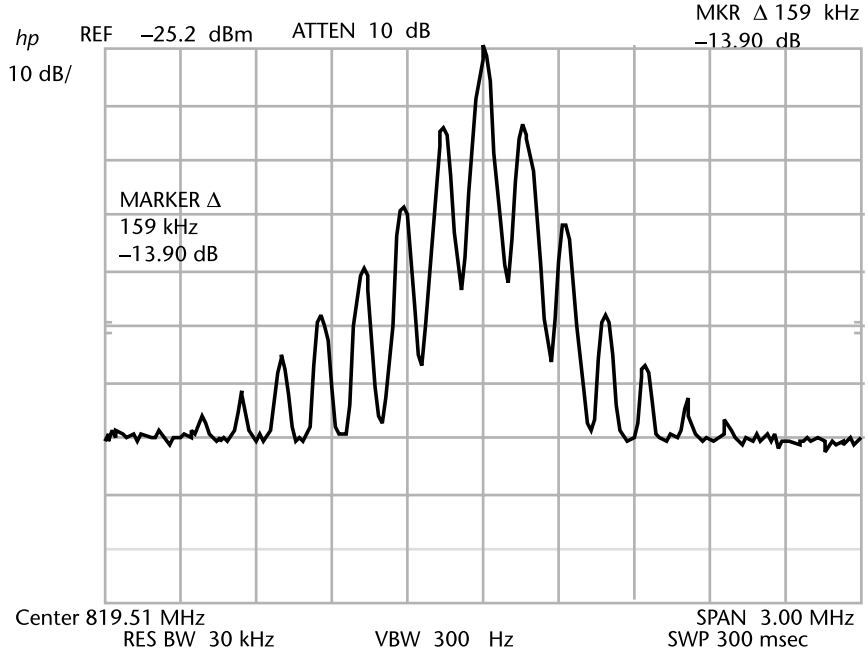

Figure 10.31 Measured output spectrum of a fractional-*N* frequency synthesizer with loop divisor $N = 160 + 1/32$ and the comparison frequency $f_r/R = 5.12$ MHz.

cycles and divided by 161 for *F-K* cycles, which results in an average division ratio of 160 + *K*/*F*. As an example, for the second channel with $K = 1$ and $F = 32$, the simulated fractional accumulator output is given in Figure 10.30. As shown, the fractional accumulator outputs a carryout in every 32 comparison cycles, which forces the loop divider to divide by 160 for 31 cycles and to divide by 161 for 1 cycle (Figure 10.30(a)). The periodic phase correction pulse due to dividing by 161 generates fractional spurs with uniform spacing of $f_r/R/32$=160 kHz as shown in Figure 10.30(b).

The spurious tones generated by the accumulator will appear in the output spectrum of the synthesizer. Figure 10.31 presents a measured spectrum at a fractional-*N* synthesizer output with loop divisor $N = 160 + 1/32$ and the comparison frequency $f_r/R = 5.12$ MHz. Fractional spurs at integer multiples of $\frac{f_r}{R} \times \frac{K}{F}$ =160 kHz are observed. The roll-off of the spur magnitude as frequency increases is due to the loop lowpass roll-off characteristics.

A fractional-*N* synthesizer achieves fine step size and low in-band phase noise with the penalty of fractional spurious tones, which comes from the periodic division ratio variation. Fractional spurs may be removable by using a high-order loop transfer function if the closest spur is outside of the PLL bandwidth. Note that the spacing of the closest spur to the carrier is determined by the synthesizer step size. For a synthesizer with fine step size smaller than the transfer function bandwidth, it is thus practically impossible to remove fractional spurs by using a loop LPF. Reducing the loop bandwidth to combat the fractional spurs means that you have to pay the penalty of longer lock time and increased out-of-band phase noise due to the VCO, as discussed earlier. Even if the closest spur is outside of the loop filter bandwidth, removing those spurs normally requires a high-order loop filter with sharp roll-off, which increases the complexity and cost of the synthesizer. To remove the fractional spurious components for a synthesizer with fine step size, the best solution is to employ a noise shaper in the fractional accumulator [2]. Their function is to break up the repeated patterns of the loop divisor time sequence without affecting its average division ratio. This will result in the reduction or elimination of the spurs in spectrum.

References

[1] Larson, L. E., (ed.), *RF and Microwave Circuit Design for Wireless Communications*, Norwood, MA: Artech House, 1997.

[2] Rogers, J., C. Plett, and F. Dai, *Integrated Circuit Design for High-Speed Frequency Synthesis*, Norwood, MA: Artech House, 2006.

[3] Craninckx, J., and M. Steyaert, *Wireless CMOS Frequency Synthesizer Design*, Dordrecht, the Netherlands: Kluwer Academic Publishers, 1998.

[4] Kroupa, V. F., "Low-Noise Microwave Synthesizer Design Principles," in *Direct Digital Frequency Synthesizers*, New York: IEEE Press, 1999, pp. 175–180.

[5] Manassewitch, V., *Frequency Synthesizers Theory and Design*, New York: John Wiley & Sons, 2005.

[6] Noordanus, J., "Frequency Synthesizers—A Survey of Techniques," *IEEE Trans. on Comm. Tech.*, Vol. 17, No. 2, April 1969, pp. 257–271.

[7] Reinhardt, V. S., et al., "A Short Survey of Frequency Synthesizer Techniques," *Frequency Control Symposium*, May 1986, pp. 355–365.

[8] Rohde, U. L., *Digital Frequency Synthesizers: Theory and Design*, Upper Saddle River, NJ: Prentice-Hall, 1983.

[9] Rohde, U. L., *Microwave and Wireless Synthesizers: Theory and Design*, New York: John Wiley & Sons, 1997.

[10] Lindsay, W. C., *Phase-Locked Loops*, New York: IEEE Press, 1986.

[11] Razavi, B., *Design of Integrated Circuits for Optical Communications*, New York: McGraw-Hill, 2002.

[12] Razavi, B., *Monolithic Phase-Locked Loops and Clock Recovery Circuits*, New York: Wiley-IEEE Press, 1996.

[13] Best, R. E., *Phase-Locked Loops: Theory, Design, and Applications*, 6th ed., New York: McGraw-Hill, 2007.

[14] Blanchard, A., *Phase-Locked Loops: Applications to Coherent Receiver Design*, New York: John Wiley & Sons, 1976.

[15] Gardner, F. M., *Phaselock Techniques*, New York: John Wiley & Sons, 1979.

[16] Egan, W. F., *Frequency Synthesis by Phase Lock*, New York: John Wiley & Sons, 2000.

[17] Kroupa, V. F., "Noise Properties of PLL Systems," *IEEE Trans. on Communications*, Vol. 30, October 1982, pp. 2244–2252.

[18] Wolaver, D. H., *Phase-Locked Loop Circuit Design*, Upper Saddle River, NJ: Prentice-Hall, 1991.

[19] Crawford, J. A., *Frequency Synthesizer Design Handbook*, Norwood, MA: Artech House, 1994.

[20] Lee, H., et al., "A - Fractional-N Frequency Synthesizer Using a Wide-Band Integrated VCO and a Fast AFC Technique for GSM/GPRS/WCDMA Applications," *IEEE J. Solid-State Circuits*, Vol. 39, July 2004, pp. 1164–1169.

[21] Leung, G., and H. Luong, "A 1-V 5.2 GHz CMOS Synthesizer for WLAN Applications," *IEEE J. Solid-State Circuits*, Vol. 39, November 2004, pp. 1873–1882.

[22] Rhee, W., B. Song, and A. Ali, "A 1.1-GHz CMOS Fractional-N Frequency Synthesizer with a 3-b Third-Order Modulator," *IEEE J. Solid-State Circuits*, Vol. 35, October 2000, pp. 1453–1460.

[23] Lo, C., and H. Luong, "A 1.5-V 900-MHz Monolithic CMOS Fast-Switching Frequency Synthesizer for Wireless Applications," *IEEE J. Solid-State Circuits*, Vol. 37, April 2002, pp. 459–470.

[24] Ahola, R., and K. Halonen, "A 1.76-GHz 22.6mW Fractional-N Frequency Synthesizer," *IEEE J. Solid-State Circuits*, Vol. 38, January 2003, pp. 138–140.

[25] Heng, C., and B. Song, "A 1.8-GHz CMOS Fractional-N Frequency Synthesizer with Randomizer Multiphase VCO," *IEEE J. Solid-State Circuits*, Vol. 38, June 2003, pp. 848–854.

[26] Park, C., O. Kim, and B. Kim, "A 1.8-GHz Self-Calibrated Phase-Locked Loop with Precise I/Q Matching," *IEEE J. Solid-State Circuits*, Vol. 36, May 2001, pp. 777–783.

[27] Klepser, B., M. Scholz, and E. Götz, "A 10-GHz SiGe BiCMOS Phase-Locked-Loop Frequency Synthesizer," *IEEE J. Solid-State Circuits*, Vol. 37, March 2002, pp. 328–335.

[28] Pellerano, S., et al., "A 13.5-mW 5-GHz Frequency Synthesizer with Dynamic-Logic Frequency Divider," *IEEE J. Solid-State Circuits*, Vol. 39, February 2004, pp. 378–383.

[29] Leenaerts, D., et al., "A 15-mW Fully Integrated I/Q Synthesizer for Bluetooth in 0.18-μm CMOS," *IEEE J. Solid-State Circuits*, Vol. 38, July 2003, pp. 1155–1162.

[30] Kan, T., G. Leung, and H. Luong, "A 2-V 1.8-GHz Fully Integrated CMOS Dual-Loop Frequency Synthesizer," *IEEE J. Solid-State Circuits*, Vol. 37, August 2002, pp. 1012–1020.

[31] Aytur, T., and B. Razavi, "A 2-GHz, 6-mW BiCMOS Frequency Synthesizer," *IEEE J. Solid-State Circuits*, Vol. 30, December 1995, pp. 1457–1462.

[32] Chen, W., et al., "A 2-V 2.3/4.6-GHz Dual-Band Frequency Synthesizer in 0.35-μm Digital CMOS Process," *IEEE J. Solid-State Circuits*, Vol. 39, January 2004, pp. 234–237.

[33] Yan, W., and H. Luong, "A 2-V 900-MHz CMOS Dual-Loop Frequency Synthesizer for GSM Receivers," *IEEE J. Solid-State Circuits*, Vol. 36, February 2001, pp. 204–216.

[34] Shu, K., et al., "A 2.4-GHz Monolithic Fractional-N Frequency Synthesizer with Robust Phase-Switching Prescaler and Loop Capacitance Multiplier," *IEEE J. Solid-State Circuits*, Vol. 38, June 2003, pp. 866–874.

[35] Shu, Z., K. Lee, and B. Leung, "A 2.4-GHz Ring-Oscillator-Based CMOS Frequency Synthesizer with a Fractional Divider Dual-PLL Architecture," *IEEE J. Solid-State Circuits*, Vol. 39, March 2004, pp. 452–462.

[36] McMahill, D., and C. Sodini, "A 2.5-Mb/s GFSK 5.0-Mb/s 4-FSK Automatically Calibrated - Frequency Synthesizer," *IEEE J. Solid-State Circuits*, Vol. 37, January 2002, pp. 18–26.

[37] Lam, C., and B. Razavi, "A 2.6-GHz/5.2-GHz Frequency Synthesizer in 0.4-μm CMOS Technology," *IEEE J. Solid-State Circuits*, Vol. 35, May 2000, pp. 788–794.

[38] Perrott, M., T. Tewksbury, and C. Sodini, "A 27-mW CMOS Fractional-N Synthesizer Using Digital Compensation for 2.5-Mb/s GFSK Modulation," *IEEE J. Solid-State Circuits*, Vol. 32, December 1997, pp. 2048–2060.

[39] Temporiti, E., et al., "A 700-kHz Bandwidth Fractional Synthesizer with Spurs Compensation and Linearization Techniques for WCDMA Applications," *IEEE J. Solid-State Circuits*, Vol. 39, September 2004, pp. 1446–1454.

[40] Dehng, G., et al., "A 900-MHz 1-V CMOS Frequency Synthesizer," *IEEE J. Solid-State Circuits*, Vol. 35, August 2000, pp. 1211–1214.

[41] Lin, T., and W. Kaiser, "A 900-MHz 2.5-mA CMOS Frequency Synthesizer with an Automatic SC Tuning Loop," *IEEE J. Solid-State Circuits*, Vol. 36, March 2001, pp. 424–431.

[42] Rategh, H., H. Samavati, and T. H. Lee, "A CMOS Frequency Synthesizer with an Injection-Locked Frequency Divider for a 5-GHz Wireless LAN Receiver," *IEEE J. Solid-State Circuits*, Vol. 35, May 2000, pp. 780–787.

[43] Hwang, I., S. Song, and S. Kim, "A Digitally Controlled Phase-Locked Loop with a Digital Phase-Frequency Detector for Fast Acquisition," *IEEE J. Solid-State Circuits*, Vol. 36, October 2001, pp. 1574–1581.

[44] Zhang, B., P. Allen, and J. Huard, "A Fast Switching PLL Frequency Synthesizer with an On-Chip Passive Discrete-Time Loop Filter in 0.25-μm CMOS," *IEEE J. Solid-State Circuits*, Vol. 38, October 2003, pp. 855–865.

[45] Da Dalt, N., et al., "A Fully Integrated 2.4-GHz LC-VCO Frequency Synthesizer with 3-ps Jitter in 0.18-μm Standard Digital CMOS Copper Technology," *IEEE J. Solid-State Circuits*, Vol. 37, July 2002, pp. 959–962.

[46] Craninckx, J., and M. S. J. Steyaert, "A Fully Integrated CMOS DCS-1800 Frequency Synthesizer," *IEEE J. Solid-State Circuits*, Vol. 33, December 1998, pp. 2054–2065.

[47] Koo, Y., et al., "A Fully Integrated CMOS Frequency Synthesizer with Charge-Averaging Charge Pump and Dual-Path Loop Filter for PCS- and Cellular-CDMA Wireless Systems," *IEEE J. Solid-State Circuits*, Vol. 37, May 2002, pp. 536–542.

[48] Bax, W. T., and M. A. Copeland, "A GMSK Modulator Using a Frequency Discriminator-Based Synthesizer," *IEEE J. Solid-State Circuits*, Vol. 36, August 2001, pp. 1218–1227.

[49] Bietti, I. et al., "An UMTS Fractional Synthesizer with 200kHz Bandwidth and -128dBc/Hz @ 1 MHz Using Spurs Compensation and Linearization Techniques," *Proc. IEEE Custom Integrated Circuits Conference*, 2003, pp. 463–466.

[50] De Muer, B. and M. Steyaert, "A CMOS Monolithic -Controlled Fractional-N Frequency Synthesizer for DCS-1800," *IEEE J. Solid-State Circuits*, Vol. 37, July 2002, pp. 835–844.

[51] Riley, T. A., M. Copeland, and T. Kwasniewski, "Delta–Sigma Modulation in Fractional-N Frequency Synthesis," *IEEE J. Solid-State Circuits*, Vol. 28, May 1993, pp. 553–559.

[52] Rogers, J. W. M., et al., "A Fractional-N Frequency Synthesizer with Multi-Band PMOS VCOs for 2.4 and 5 GHz WLAN Applications," *European Solid-State Circuits Conference (ESSCIRC)*, September, 2003, pp. 651–654.

[53] Vaucher, C. S., et al., "A Family of Low-Power Truly Modular Programmable Dividers in Standard 0.35μm CMOS Technology," *IEEE J. Solid-State Circuits*, Vol. 35, July 2000, pp. 1039–1045.

[54] Razavi, B., *RF Microelectronics*, Upper Saddle River, NJ: Prentice-Hall, 1998.

[55] Kroupa, V. F., "Jitter and Phase Noise in Frequency Dividers," *IEEE Trans. on Instrumentation and Measurement*, Vol. 50, No. 5, October 2001, p. 1241.

[56] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1984.

[57] Sze, S. M., *Physics of Semiconductor Devices*, 2nd ed., New York: John Wiley & Sons, 1981.

[58] Gray, P. R., et al., *Analysis and Design of Analog Integrated Circuits*, 4th ed., New York: John Wiley & Sons, 2001.

[59] Amaya, R. E., et al., "EM and Substrate Coupling in Silicon RFICs," *IEEE J. Solid State Circuits*, Vol. 40, No. 9, September 2005.

[60] Leeson, D. B., "A Simple Model of Feedback Oscillator Noise Spectrum," *Proc. IEEE*, February 1966, pp. 329–330.

[61] Uwano, T., et al., "Design of a Low-Phase Noise VCO for an Analog Cellular Portable Radio Application," *Electronics and Communications in Japan*, Vol.77, 1994, pp. 58–65.

[62] Dow, S., et al., "A Dual-Band Direct-Conversion Transceiver IC for GSM," *Proc. International Solid-State Circuits Conference*, 2002, pp. 230–462.

[63] Vig, J. R., *Quartz Crystal Resonators and Oscillators*, U.S. Army Electronics Technology and Devices Report, SLCET-TR-88-1, 1988.

[64] Watanabe, Y., et al., "Phase Noise Measurements in Dual-Mode SC-Cut Crystal Oscillators," *IEEE Trans. on Ultrasonics, Ferroelectrics and Frequency Control*, Vol. 47, No. 2, March 2000, pp. 374–378.

[65] Rohde, U. L., "A Novel RFIC for UHF Oscillators," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, 2000, pp. 53–56.

[66] Kivinen, J., and P. Vainikainen, "Phase Noise in a Direct Sequence Based Channel Sounder," *8th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Vol. 3, 1997, pp. 1115–1119.

CHAPTER 11

Power Amplifiers

11.1 Introduction

Power amplifiers, also known as PAs, are used in the transmit side of RF circuits, typically to drive antennas. Power amplifiers typically trade off efficiency and linearity, and this trade-off is very important in a fully monolithic implementation. Higher efficiency leads to extended battery life and this is especially important in the realization of small, portable products. There are some additional challenges specifically related to being fully integrated. Integrated circuits typically have a limited power supply voltage to avoid breakdown, as well as a metal migration limit for current. Thus, simply achieving the desired output power can be a challenge. Power amplifiers dissipate power and generate heat, which has to be removed. Due to the small size of integrated circuits, this is a challenging exercise in design and packaging. Several recent overview presentations have highlighted the special problems with achieving high efficiency and linearity in fully integrated power amplifiers [1–3].

Power amplifiers are among the last circuits to be fully integrated. In many instances, there is no choice but to design them with discrete power transistors, or at least separately from the rest of the radio frequency front-end circuits. There is a lot of interest in discrete and semi-integrated power amplifier design, and for years the main reference was the classic book by Krauss, Bostian, and Raab [4]. Only recently has this book been complemented by a number of fine new books on the topic [5–7], showing that this subject matter is still of interest and may in fact be growing in importance.

11.2 Power Capability

One of the main goals of PA design is to deliver a given power to a load. This is determined to a large degree by the load resistor and the power supply. Given a particular power supply voltage V_{CC}, such as 3V, and a load resistance R_L, such as 50 , it is possible to determine the maximum power to be

$$P = \frac{V_{CC}^2}{2R} = \frac{3^2}{2 \quad 50} = 90 \text{ mW} \qquad 19.5 \text{ dBm} \tag{11.1}$$

This assumes we have a tuned amplifier and an operating point of 3V, a peak negative swing down to 0V and a peak positive swing up to 6V. In more recent

processes, because of breakdown considerations, the fastest transistors are limited to an operating point of about 1V, limiting the maximum power to about 10 mW or 10 dBm when directly driving a 50 load. Higher power is often needed and, as we will see later, is in fact possible if the load impedance is not fixed at 50 .

11.3 Efficiency Calculations

Efficiency η, sometimes also called dc-to-RF efficiency, is the measure of how effectively power from the supply is converted into output power and is given by

$$\eta = \frac{P_{\text{out}}}{P_{\text{dc}}} \tag{11.2}$$

where P_{out} is the ac output power and is given by

$$P_{\text{out}} = \frac{i_1 v_1}{2} = \frac{i_1^2 R_L}{2} \tag{11.3}$$

where i_1 and v_1 are the peak fundamental components of the current and voltage, respectively. These are determined from the actual current and voltage by Fourier analysis. P_{dc} is the power from the supply and is given by

$$P_{\text{dc}} = \frac{1}{T}\int_0^T V_{\text{CC}} i_C dt = \frac{V_{\text{CC}}}{T}\int_0^T i_C dt = V_{\text{CC}} I_{\text{dc}} \tag{11.4}$$

where I_{dc} is the dc component of the current waveform.

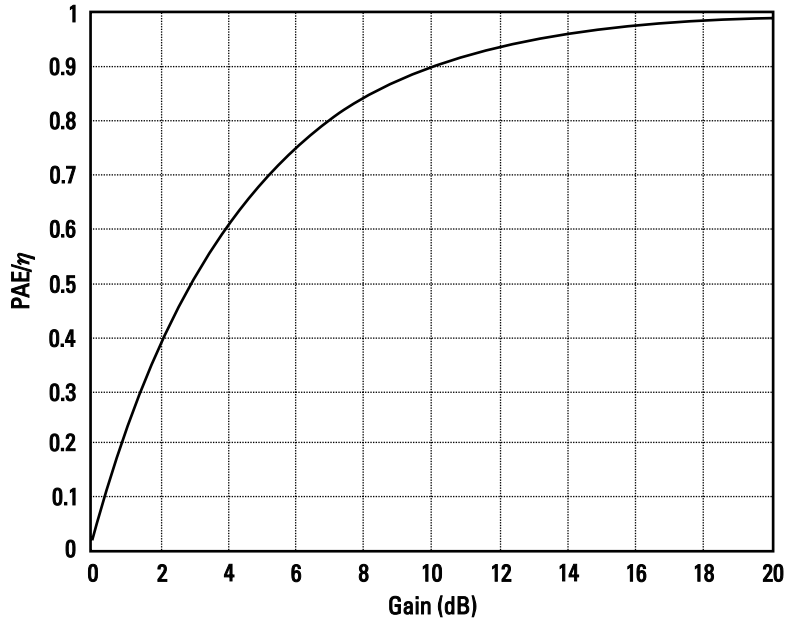

Figure 11.1 Normalized power-added efficiency versus gain.

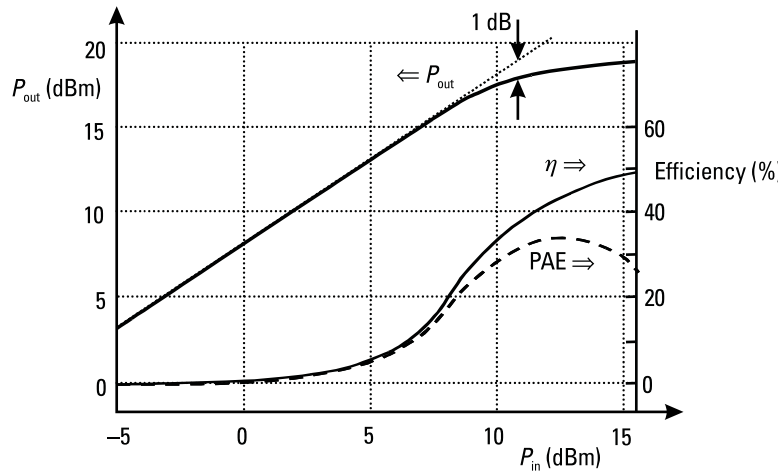

Figure 11.2 Output versus input power.

Power-added efficiency (PAE) is the same as efficiency; however, it takes the gain of the amplifier into account as follows:

$$\text{PAE} = \frac{P_{\text{out}} \quad P_{\text{in}}}{P_{\text{dc}}} = \frac{P_{\text{out}} \quad P_{\text{out}}/G}{P_{\text{dc}}} = \eta \ \ 1 \quad \frac{1}{G} \div \tag{11.5}$$

where G is the power gain $P_{\text{out}}/P_{\text{in}}$. Thus, it can be seen that for high gain, power-added efficiency, PAE is the same as dc-to-RF efficiency η.

Figure 11.1 shows the efficiency in comparison to power-added efficiency for a range of power gains. It can be seen that for gain higher than 10 dB, PAE is within 10% of the efficiency η. For lower gain, PAE is lower relative to η. For example, if the gain is 3 dB, the PAE is only half of the dc-to-RF efficiency.

A typical plot of output power and efficiency versus input power is shown in Figure 11.2. It can be seen that while efficiency keeps increasing for higher input power, as the amplifier compresses and gain decreases, the power-added efficiency also decreases. Thus, there is an optimal value of power-added efficiency and it typically occurs a few decibels beyond the 1-dB compression point.

11.4 Matching Considerations

In order to obtain maximum output power, typically the power amplifier is not conjugately matched. Instead, the load is designed such that the amplifier has the correct voltage and current to deliver the required power. We note that conjugate matching means that $_S = S_{11}{}^*$ and $_L = S_{22}{}^*$ as shown in Figure 11.3. In the figure, $_S$ is the source reflection coefficient and $_L$ is the load reflection coefficient.

11.4.1 Matching to $S_{22}{}^*$ Versus Matching to $_{\text{opt}}$

For low input power where the amplifier is linear, the maximum output power is obtained with $_L = S_{22}{}^*$. However, this value of S_{22} will not be the optimum load

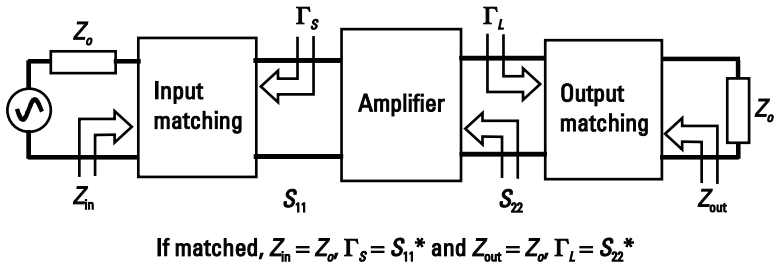

Figure 11.3 Block diagram of amplifier and matching circuits.

for high input power where the amplifier is nonlinear. Nonlinearities result in gain compression, the appearance of harmonics, and additional phase shift. The result can be a shift of the operating point and a shift in the optimal load impedance. For these reasons, for large-signal operation, tuning is done by determining the optimal load $_{opt}$, typically by doing an exhaustive search called a *load pull*. The comparison between tuning for small signal and large signal is shown in Figure 11.4.

As illustrated, the small-signal tuning curve results in higher output power for small signals, while the large-signal tuning curve results in higher output power for larger signals. Typically, if the operation is at the optimal PAE point (as shown in Figure 11.2), optimal-power tuning produces about 1 to 3 dB of higher power. Gain is reduced (for small P_{in}) typically by a slightly smaller amount.

An estimate of the optimum impedance $_{opt}$ can be made by adjusting the load so that the transistor current and voltage go through their maximum excursion, as shown in Figure 11.5, with the output susceptance (typically capacitive susceptance) reactively matched.

11.5 Class A, B, and C Amplifiers

Power amplifiers are grouped into classes depending on the nature of their voltage and current waveforms. The first major classes to be considered are class A, B, and C amplifiers. Figure 11.6 shows a simple amplifier circuit for both bipolar and CMOS that can be used for any of these classes. Waveforms for the base voltage v_B, collector voltage v_C. and collector current i_C, are shown in Figure 11.7 for class A

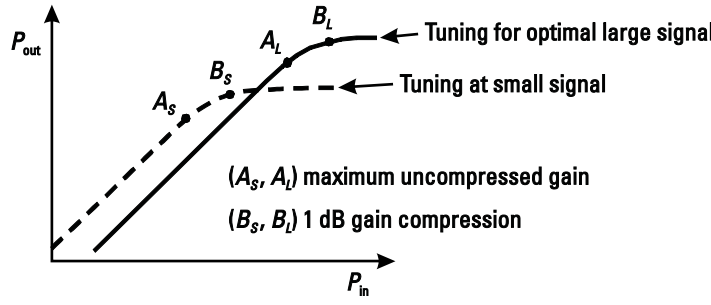

Figure 11.4 Optimal matching versus small-signal matching.

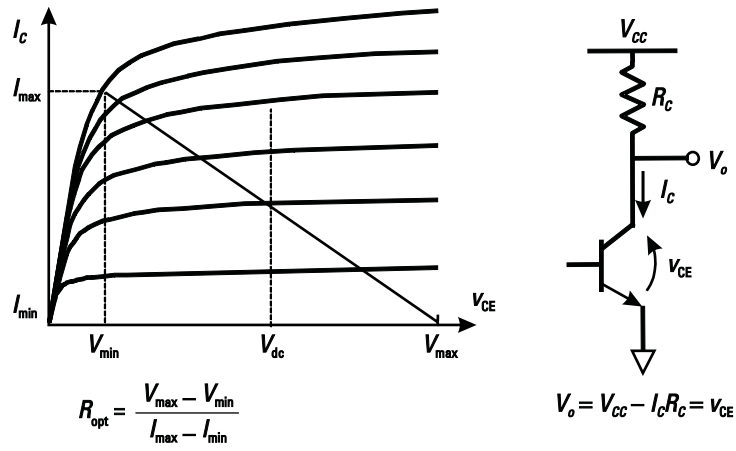

Figure 11.5 Current and voltage excursion of power amplifier.

operation and in Figure 11.8 for class B and C operation. For CMOS-based amplifiers, the input would be on the gate and the output on the drain. Class A amplifiers can be designed to have more gain than class B or class C amplifiers. However, as will be seen later, the achievable output power is nearly the same for class A, class AB, or class B amplifiers. For a class C amplifier, where the transistor conducts for a short part of the period, the output power is reduced.

The maximum sinusoidal collector voltage is shown from approximately 0V to $2V_{CC}$. The assumption that the collector voltage can swing down close to 0V simplifies the analysis, but results in output power and efficiency being somewhat overestimated. While the collector voltage is assumed to be sinusoidal because of the filtering action of the tuned circuit, the collector current may be sinusoidal, as in class A operation, or may be nonsinusoidal, as in class B or C operation, which is determined mainly by how the transistor is biased.

The classifications as A, AB, B, or C describe the fraction of the full cycle for which current is flowing in the driver transistor. Such a fraction can be described as conduction angle, which is the number of degrees (out of 360°) for which current

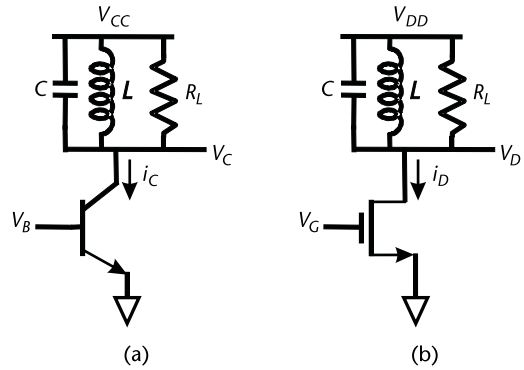

Figure 11.6 Power amplifier circuits with tuned load: (a) bipolar implementation and (b) CMOS implementation.

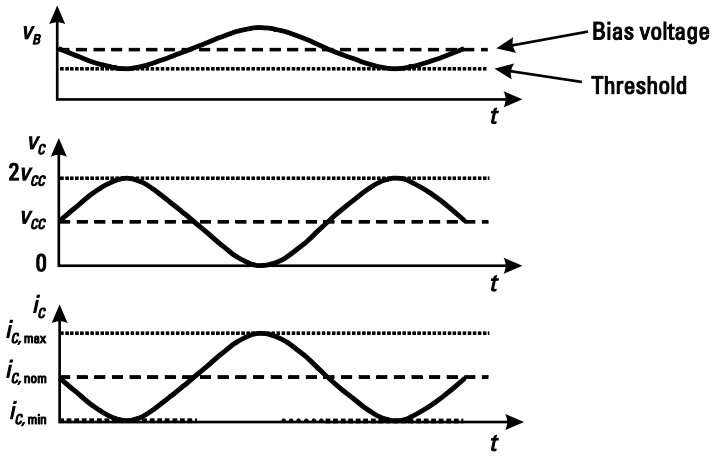

Figure 11.7 Waveforms for a class A power amplifier.

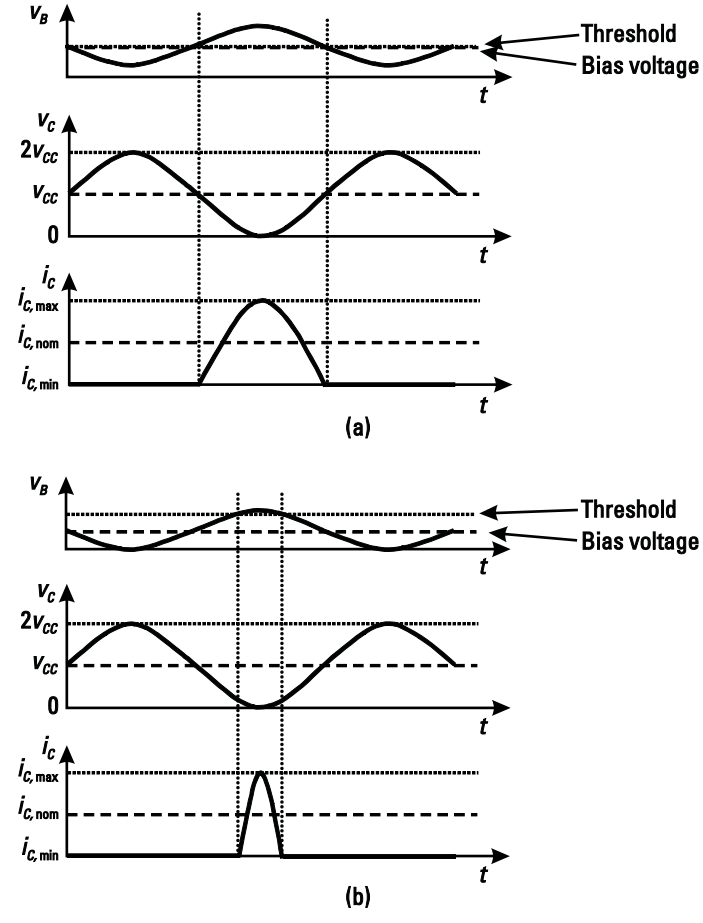

Figure 11.8 Power amplifier waveforms: (a) class B operation and (b) class C operation.

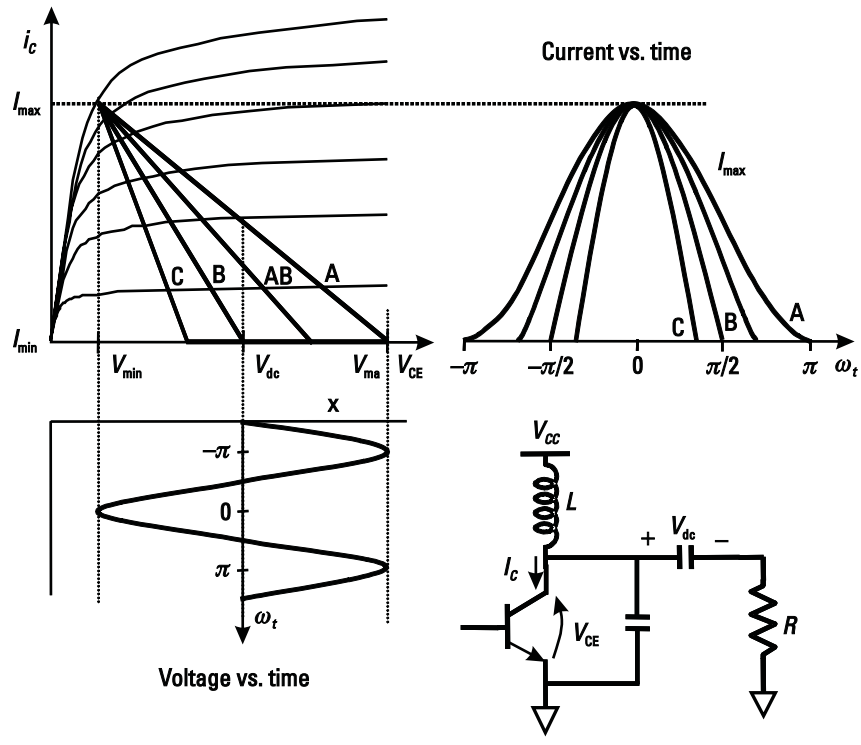

Figure 11.9 Current and voltage excursions for different classes of amplifiers.

is flowing. If current is always flowing, the conduction angle is 360° and operation is class A. If the current flows for exactly half of the time, the conduction angle is 180° and operation is class B. For conduction angles between 180° and 360°, the operation is class AB. If the current flows between 0° and 180°, the transistor is said to be operating as class C. It will be shown later that a higher conduction angle will result in better linearity but lower efficiency. A summary is shown in Table 11.1.

To obtain a high efficiency, power loss in the transistor must be minimized, and this means that the current should be minimum while voltage is high, and voltage should be minimum while the current is high. It can be seen in Figures 11.7 and 11.8 that for all waveforms, the maximum voltage is aligned with the minimum current, and the maximum current is aligned with minimum voltage. It can further be seen that for class B and class C, the current is set to zero for part of the cycle where the voltage is high. This leads to increased efficiency; however, there will still

Table 11.1 PA Class A, AB, B, and C Conduction Angle and Efficiency

Class	*Conduction Angle (°)*	*Efficiency (Maximum Theoretical)(%)*	*Ouput Power (Normalized)*
A	360	50	1
AB	360–180	50–78.5	Nearly constant about 1 (theory: maximum of 1.07 at 245°)
B	180	78.5	1
C	180–0	78.5–100	1 at 180°, 0 at 0°

be some loss, since there is an overlap of nonzero voltage and current. Other classes of amplifiers, to be described in later sections, namely, classes D, E, and F, are designed such that the voltage across the transistor is also nonlinear, leading to higher efficiencies, in some cases up to 100%. A different way to improve efficiency, while potentially maintaining linearity, is to power a linear amplifier from a variable or switched power supply. This is the basis for class G and H designs. All of the above amplifiers will be discussed in more detail in Sections 11.6 to 11.9.

Figure 11.9 shows a simplified power amplifier and a plot of transistor current versus time for the various classes. The different classes can be obtained with the same circuit by adjusting the input bias circuit. For example, in class A, if the maximum current is I_{max}, the amplifier is set to have a nominal bias of half of I_{max} so that current swings from nearly 0 to I_{max}. For class B, the bias is set so that the transistor is nominally at the edge of conduction so that the positive input swing will cause the transistor to conduct, while negative input will guarantee the transistor is off. Thus, the transistor will conduct half the time.

Class A, B, C Analysis

Except for class A, the current through the transistor is not sinusoidal, but may be modeled as a biased sinusoid as shown in Figure 11.10.

The collector current (or drain current for CMOS) can be expressed as

$$i_C = I_{CC}\cos\omega t \quad I_{CQ} \tag{11.6}$$

This is valid from $\theta < \omega t < \theta$. Here, I_{CQ} is given by

$$I_{CQ} = I_{CC}\cos\theta \tag{11.7}$$

We can find the dc component and the fundamental component by determining the Fourier series of this waveform. Note that the tuned circuit will give us the fundamental component (if tuned to f_0).

$$I_{dc} = \frac{1}{2\pi}\int_{\theta}^{\theta}[I_{CC}\cos\omega t \quad I_{CQ}]d(\omega t) = \frac{1}{\pi}\int_{0}^{\theta}[I_{CC}\cos\omega t \quad \cos\theta]d(\omega t)$$

$$= \frac{I_{CC}}{\pi}[\sin\theta \quad \theta\cos\theta] \tag{11.8}$$

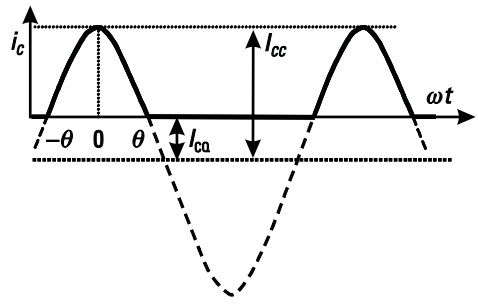

Figure 11.10 Waveform in analysis of class A, B, and C amplifiers.

Power supplied is given by

$$P_{CC} = V_{CC} I_{dc} = \frac{V_{CC} I_{CC}}{\pi}(\sin\theta \quad \theta\cos\theta) \tag{11.9}$$

The fundamental current i_1 given by

$$i_1 = \frac{4}{2\pi}\int_0^{\theta}(I_{CC}\cos\omega t \quad I_{CQ})\cos\omega t d(\omega t) = \frac{I_{CC}}{2\pi}[2\theta \quad \sin 2\theta] \tag{11.10}$$

Output power is given by

$$P_{out} = \frac{i_1^2 R_L}{2} = \frac{v_{peak}}{\sqrt{2}} \times \frac{i_{peak}}{\sqrt{2}} \tag{11.11}$$

The maximum possible v_{peak} is when the output swings from about 0V to $2V_{CC}$ or $v_{peak} = V_{CC.}$ Thus,

$$P_{out,max} = \frac{V_{CC}}{\sqrt{2}} \times \frac{i_1}{\sqrt{2}} = \frac{V_{CC} I_{CC}}{4\pi}[2\theta \quad \sin 2\theta] \tag{11.12}$$

Efficiency for this maximum possible voltage swing is given by

$$\eta_{max} = \frac{P_{out,max}}{P_{dc}} = \frac{2\theta \quad \sin 2\theta}{4(\sin\theta \quad \theta\cos\theta)} \tag{11.13}$$

The efficiency is plotted in Figure 11.11.

The actual output power for an output peak voltage of V_{op} can be found as a function of θ:

$$P_{out} = \frac{V_{op} I_{CC}}{4\pi}[2\theta \quad \sin 2\theta] \tag{11.14}$$

noting that to get maximum power, the load resistance has to be adjusted so that the maximum voltage $v_{o,max}$ is approximately $2V_{CC}$ and the minimum voltage $v_{o,min}$ is approximately zero; thus, V_{op} is equal to V_{CC}.

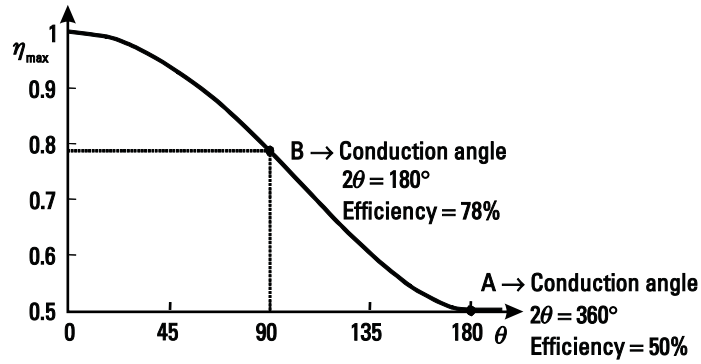

Figure 11.11 Maximum efficiency versus conduction angle.

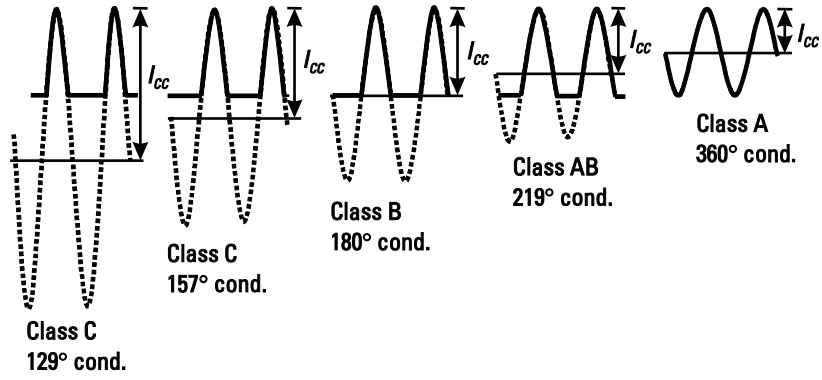

Figure 11.12 Time-domain waveforms for various conduction angles.

Equation (11.14) will be more convenient to plot if we eliminate I_{CC}. Recall that I_{CC} is a measure of the peak current, not with respect to zero, but with respect to the center of the sine wave where the center for class C is less than zero, as shown in Figure 11.12.

The peak current is

$$i_{\text{peak}} = I_{\text{CC}}(1 \quad \cos\theta) \tag{11.15}$$

For the same peak current, the maximum distorted, or class A, output power can be determined by noting that the voltage goes from 0 to $2V_{\text{CC}}$ and the current goes from 0 to i_{peak}. Then converting peak-to-peak to rms, we obtain normalized output power $P_{\text{o,norm}}$ as

$$P_{o,\text{norm}} = \frac{(I_{\max} \quad I_{\min})(V_{\max} \quad V_{\min})}{8} = \frac{I_{\text{CC}}V_{\text{CC}}(1 \quad \cos\theta)}{4} \tag{11.16}$$

Then we could plot normalized, maximum output power as

$$P_{o,\max,\text{norm}} = \frac{P_{o,\max}}{P_{o,\text{norm}}} = \frac{1}{\pi}\frac{(2\theta \quad \sin 2\theta)}{1 \quad \cos\theta} \tag{11.17}$$

This has been plotted in Figure 11.13. It can be seen that at a conduction angle of 180° and 360° (or $\theta = 90°$ and 180°), the normalized maximum output power is 1. In between is a peak with a value of about 1.07 at a conduction angle of about 245°. For maximum output power, this might appear to be the optimum conduction angle; however, it can be noted that in real life, or in simulations with other models for the current (rather than the tip of a sinusoid), this peak does not occur. However, overall, Figure 11.13 is a fairly good description of real life performance.

As an example, if $V_{\text{CC}} = 3$V, then $V_{\max}$ can go to 6V. If $I_{\max}$ is 1A, then the maximum output power is given by $P_{\text{out,max}} = (V_{\max} \bullet I_{\max})/8 = 0.75$W at $\theta = 90°$ and at $\theta = 180°$.

An expression that is valid for $n = 2$ or higher can be found for i_n, the peak current of the nth harmonic:

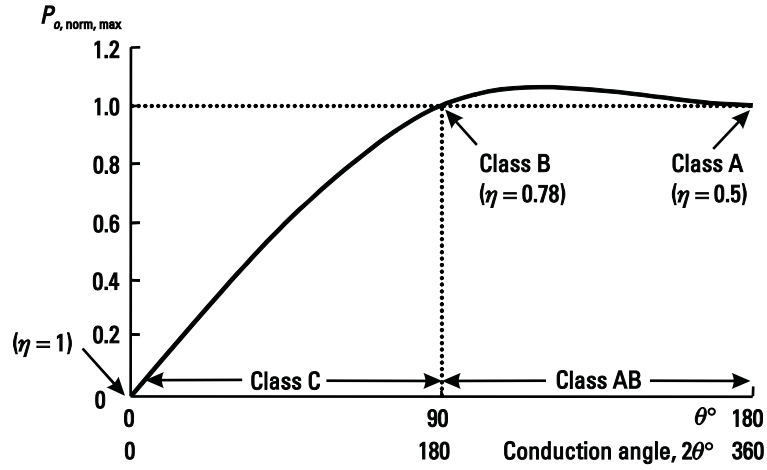

Figure 11.13 Maximum output power versus conduction angle.

$$i_n = \frac{2I_{CC}}{\pi} \quad \frac{\cos\theta\sin n\theta \quad n\sin\theta\cos n\theta}{n(n^2 \quad 1)} \tag{11.18}$$

Figure 11.14 shows the current components normalized to the maximum current excursions in the transistor. The dc component is found from (11.8), the fundamental component is found from (11.10), and the other components are found from (11.18), with normalization done using (11.15).

We note that for class A ($\theta = 180°$ or conduction angle is 360°) the collector current is perfectly sinusoidal and there are no harmonics. At lower conduction

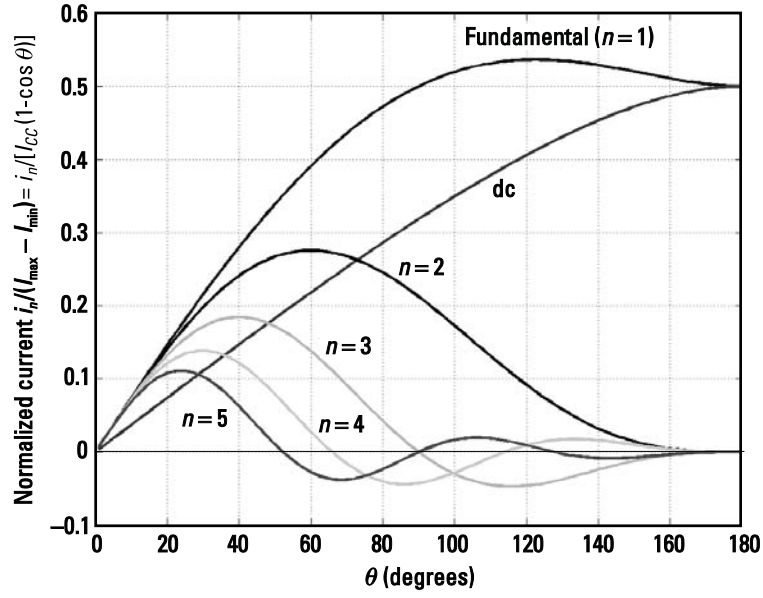

Figure 11.14 Fourier coefficients for constant transconductance.

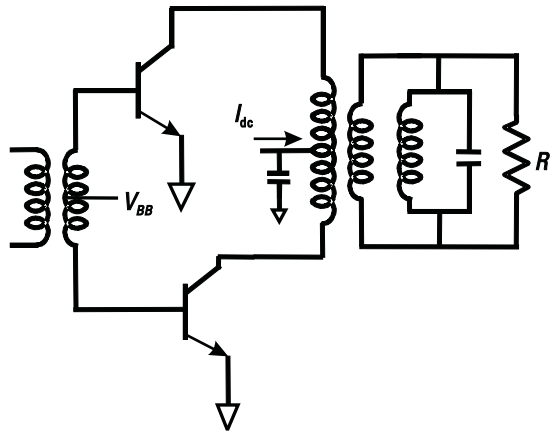

Figure 11.15 Push-pull class B amplifier.

angles, the collector current is rich in harmonics. However, the tuned circuit load will filter out most of these, leaving only the fundamental to make it through to the output.

At very low conduction angles, the current "pulse" is very narrow approaching the form of an impulse in which all harmonic components are of equal amplitude. Here efficiency can be high, but the output power is lower.

11.5.1 Class B Push-Pull Arrangements

In the push-pull arrangement shown with bipolar transistors in Figure 11.15 with transformers or with CMOS transistors in Figure 11.16 with power combiners, each transistor is on for half the time. Thus, the two are on for the full time resulting in the possibility of low distortion yet with class B efficiency, with a theoretical maximum of 78%. The total output power is twice that of each individual transistor.

Mathematically, each transistor current waveform as shown in Figure 11.16 is described by the fundamental and the even harmonics as shown in (11.19).

$$i_A = \frac{I_P}{\pi} + \frac{I_P}{2}\cos\omega t + \frac{2I_P}{3\pi}\cos 2\omega t \quad \frac{2I_P}{15\pi}\cos 4\omega t + \ldots \tag{11.19}$$

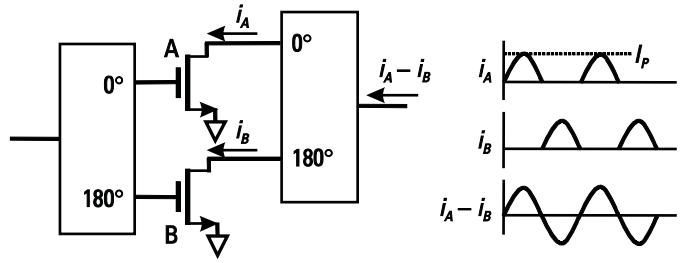

Figure 11.16 Class B amplifier with combiner.

Note that this agrees with Figure 11.14, which showed the third and fifth harmonics passing through 0 when $\theta = 90°$ (conduction angle is 180°). In the push-pull arrangement, the dc components and even harmonics cancel, but odd harmonics add, thus the output contains the fundamental only, as shown in:

$$i_A \quad i_B \quad I_P \cos \omega t \tag{11.20}$$

Note that this ideal output of only the fundamental component is only valid if the amplifier is not driven hard.

11.5.2 Models for Transconductance

There is an error in assuming current is the tip of a sine wave for a sinusoidal input. This model assumes that $i = g_m v_{\text{in}}$ with a constant value for g_m as long as v_{in} is larger than the threshold voltage. There are at least two other more realistic models for g_m. One is to assume that g_m is linearly related to the input voltage, as might be the case for a bipolar amplifier with emitter degeneration or a CMOS amplifier with source degeneration. Another model relates the transconductance g_m exponentially to the input voltage. While these models are more realistic, only the model with constant g_m results in easy analytical equations, but numerical results can be obtained for all cases. The resulting output powers and efficiencies are similar to the results for the simple constant g_m assumption (typically with somewhat reduced output power and efficiency). However, for high frequency, none of these models are completely accurate, so it is recommended that the simple constant g_m model be used for speedy hand calculations and full simulations be used to continue the design.

Example 11.1: Class A Power Amplifier

Design a class A power amplifier that will drive 200 mW into a 50 load at 1 GHz from a 3-V power supply. The bipolar transistor unit cell that is available has the f_T versus current relationship shown in Figure 11.17. Use as many of these in parallel as necessary.

Solution:
Assuming the output voltage is centered at 3V and has a peak swing of 2.6V, the output resistance can be determined.

$$P_o = \frac{v_{\text{rms}}^2}{R} \quad \text{or} \quad R = \frac{v_{\text{rms}}^2}{P_o} = \frac{2.6^2}{2 \quad 0.2} = 16.9$$

Thus, we estimate that we will need a 16.9 load resistance for 200 mW of output power. This means we will need a matching circuit to convert between 16.9 and 50 .

We can also determine the current.

$$P_o = \frac{v_p \quad i_p}{2}$$

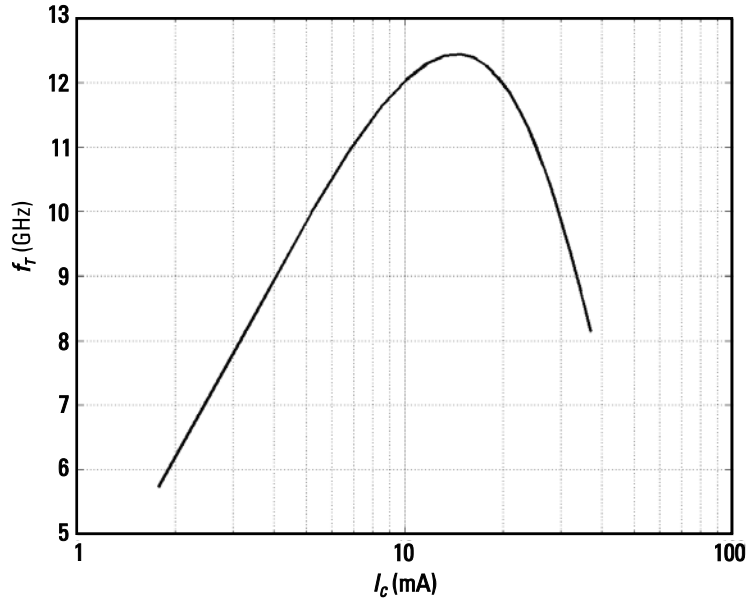

Figure 11.17 Power transistor f_T versus current.

or

$$i_p = \frac{2P_o}{v_p} = \frac{2 \quad 0.200}{2.6} = 0.154 \text{ or } 154 \text{ mA}$$

Thus, for class A, the nominal current should be about 154 mA with a peak excursion from about 0 to 308 mA. Thus, with the transistor as given, with peak f_T at about 15 mA, let us choose to use 10 of these units. Then with operation close to the peak f_T, the resulting simulated collector current is 147 mA and with a dc current gain of about 85, the input current is set to about 1.75 mA.

It was noted that the circuit had a potential stability problem for low-impedance inputs. This problem was minimized with a 5 resistor in series with the input. The input impedance was then seen to be about $5.55 - j3.14$. Through simulations, the input was matched with 3 nH of inductance in series and 9 pF of capacitance in parallel. Note that the finite Q of the input series inductor will help with stabilization. Then input power was swept to determine the power level at which the negative peak of the current reached zero. This power was used as a starting point for several iterations of sweeps of load pull and input power used to determine the optimal output load and the required input power. The transistor current crosses zero for an input power of about 8 dBm, as shown in Figure 11.18. The load pull shown in Figure 11.19 indicated that the optimal load should be $9 + j7.6$. As a series circuit, this is a little bit lower than the predicted 16.9 , indicating that the transistor could drive a bit more current than predicted. However, as a parallel circuit, this is equivalent to a parallel combination of an inductance with a 15.4 resistor, close to the predicted value. The inductive portion of

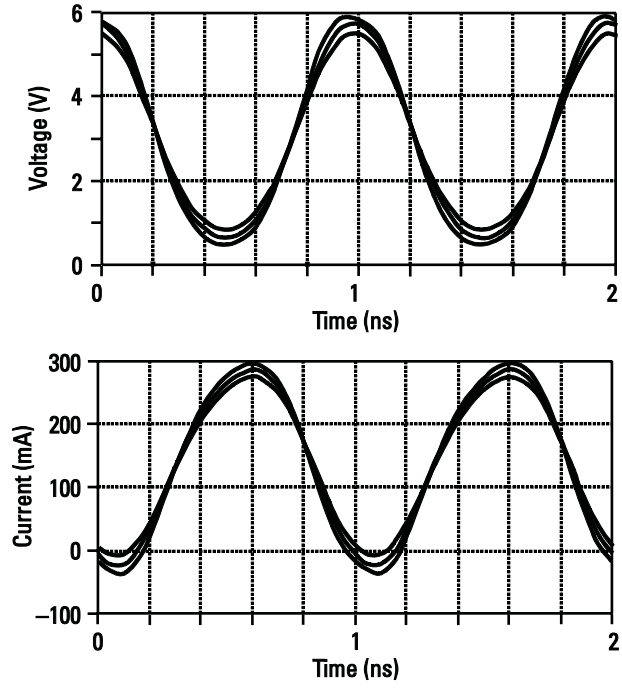

Figure 11.18 Voltage and current waveforms for input power levels of 8, 9, and 10 dBm.

the load ($j7.6$ if in series) accounts largely for the transistor output capacitance. The sweeps of P_{out} and power-added efficiency versus P_{in} shown in Figure 11.20 shows that 1-dB compression occurs at an input power of about 9 or 10 dBm and that power-added efficiency is just over 30% at an input power of 8 dBm, rising to about 42% at 10 dBm. The output power is about 23 dBm as required. The

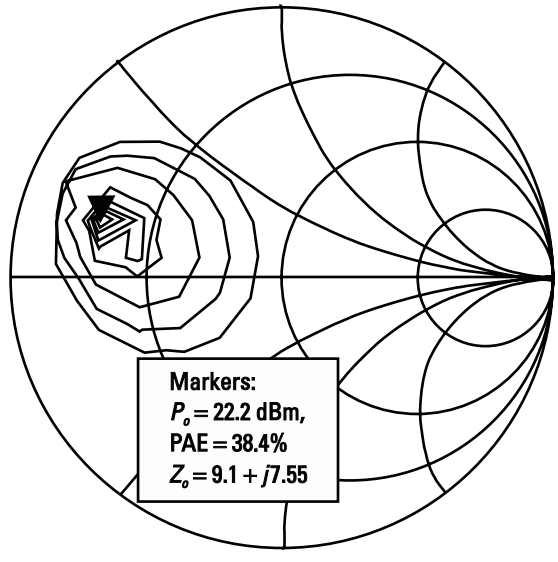

Figure 11.19 Load pull for an input power of 8.5 dBm.

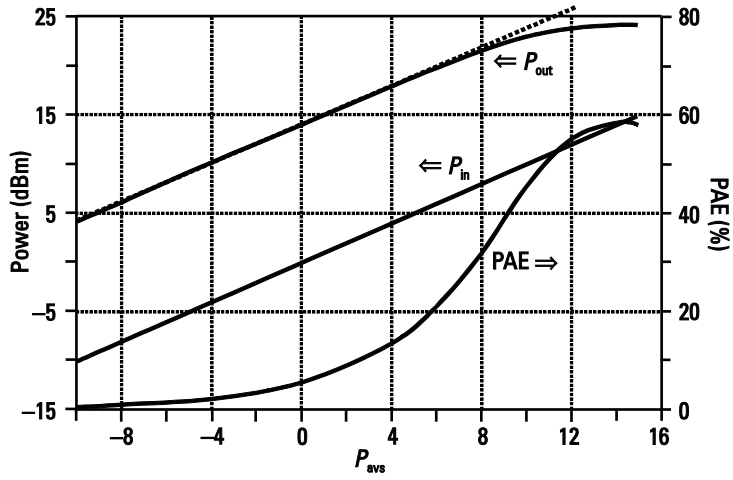

Figure 11.20 Gain and power-added efficiency versus input power.

third curve shows P_{in}, the simulated actual input power, versus P_{AVS}. If the circuit is matched, P_{in} should equal P_{AVS}. Several differences can be seen between this simulation and simple theory. The simple equations were derived assuming that output voltage swings from 0V to 6V. This does not happen, and thus power is a little bit low. This also directly leads to a lower optimal load impedance than was initially calculated. In this example, ideal models were used for passives and packaging. Obviously, the use of realistic lossy components will result in a reduction in efficiency.

Example 11.2: Bipolar Class AB Power Amplifier

As a continuation of the previous example, design a class AB power amplifier that will drive 200 mW into a 50 load at 1 GHz from a 3-V power supply. As before, the bipolar transistor unit cell that is available had the f_T versus current relationship shown in Figure 11.17. Design for the optimal PAE.

Solution:

The first thing to note is that the current extremes, $I_{\max}$ and $I_{\min}$, will still be approximately the same as they were for the class A amplifier even though current will flow for a smaller percentage of the time. Examination of Figure 11.14 shows that the output fundamental current will be roughly constant for any conduction angle between 180° and 360°. Thus, it would seem that it should be possible to reduce the nominal current through this amplifier and achieve roughly the same output power by driving the amplifier into compression. In practice, with reduced bias, the achievable output power is reduced, as shown by simulation results summarized in Table 11.2.

It can be seen that by reducing the bias current to 105 mA, approximately the same results are obtained as for the class A amplifier. However, there are a few important differences. Because the amplifier is driven into compression, the efficiency

Table 11.2 Simulation Results for Class AB Power Amplifier Example

I_{bias} (mA)	Opt PAE (%)	P_{in} (Opt PAE) (dBm)	P_{out} (Opt PAE) (dBm)	Compression (dB)	$_{opt}$ ( )
63	55.8	12	20.7	2.5	11.5 + *j*18.7
105	57.2	13	22.6	3.2	12.2 + *j*14.3
147	58.9	14	24.1	3.7	13.1 + *j*10.3

is now 57.2% and the amplifier is now nonlinear. If instead the same amplifier is used as in Example 11.1, a bias current of 147 mA, the output power is increased to 24.1 dBm and efficiency is increased to 58.9%. Time-domain waveforms for this case can be seen in Figure 11.21. It can be seen that waveforms do not match the simple theory in that voltage is not sinusoidal and the current goes negative due to transistor capacitance. By considering the positive portion of the current, the conduction angle for a 14-dBm input is estimated to be about 260° ($\theta = 130°$) and from Figure 11.11, efficiency is expected to be about 60%, which is close to the simulated value.

Example 11.3: CMOS Class AB Power Amplifier

Design a power amplifier in 130-nm CMOS that will deliver 75 mW into 50 , from 5.1 to 5.3 GHz.

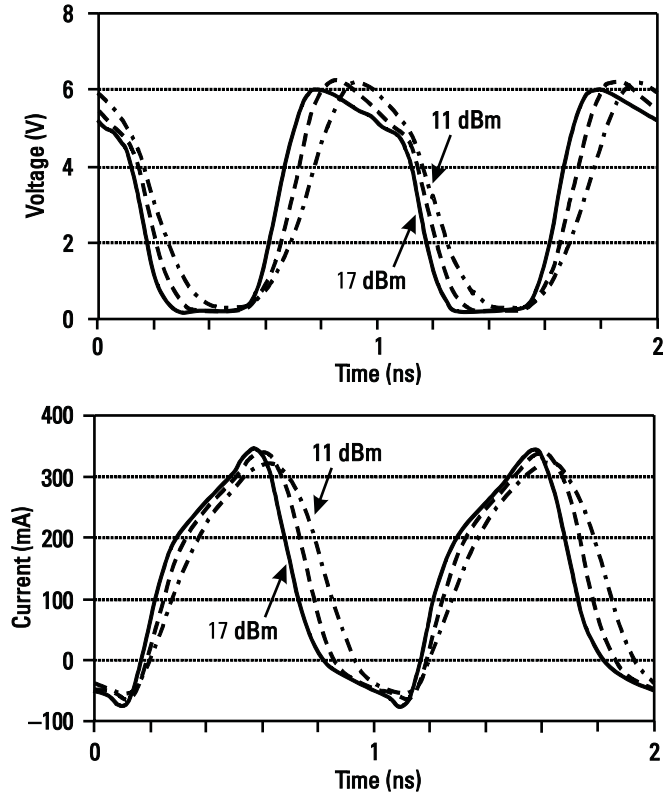

Figure 11.21 Voltage and current waveforms for bias current of 147 mA and input powers of 11, 14, and 17 dBm.

Solution:

The high-frequency transistors in the process have a suggested power supply voltage of 1.2V and an oxide breakdown voltage (V_{DG} or V_{GS}) of about 1.6V. The drain voltage will have an average value equal to the power supply voltage; hence, it will not take much positive swing before breakdown occurs. It is possible to use a cascoded structure with $V_{D2} = V_{G2} = 1.2V$. To achieve the required power might result in the output swing going up to close to 3V, meaning that the cascode transistor would have a V_{DG} of 1.8V and oxide breakdown would occur. In order to reduce the likelihood of breakdown, it is possible to limit the output voltage swing, for example, by using a smaller load resistance, to reduce the power supply voltage somewhat, to use a feedback system to have the cascode voltage follow the output voltage, or to use higher breakdown transistors. The latter approach is followed here, using a thicker oxide 3.3-V transistor with an oxide breakdown of about 3.6V. The disadvantage is that the minimum channel length is about three times higher and the transistors are slower. Curves of effective f_{max} versus bias current for three different power supply voltages are shown in Figure 11.22(a) for a cascoded minimum length transistor with a total gate width of 1 mm. Figure 11.22(b) shows f_{max} at 100 mA as a function of power supply voltages for the same cascode pair showing that V_{DD} needs to be at least 1.7V as f_{max} drops dramatically for lower voltages. The reason for this is that we have kept the current at 100 mA. To operate at lower voltages, the current density would need to be dropped. For example, at 1.2V, optimum current is only 40 mA, making it difficult to achieve the required output power level. Instead, if a 1.2-V power supply is mandatory, it would be possible to use a single transistor, but the result is low f_T, low gain, and hence low PAE. As well, single transistors have less reverse isolation and reduced stability. For this example, we will use a cascode structure with a V_{DD} of 1.8V.

From Figure 11.22, with a V_{DD} of 1.8V, the current can swing up to about a 200-mA peak. This was checked with transistor characteristics as in Figure 11.23.

From the curve, assuming the current swing is from 0 to 200 mA, a class A amplifier would be biased at 100 mA with nominal V_{DS} of V_{DD} or 1.8V. If V_{GS} can go to 2V, 200 mA occurs at V_{DS} of about 0.3V, and hence the negative swing is 1.5V. Assuming the voltage swing is symmetrical, the highest voltage is about

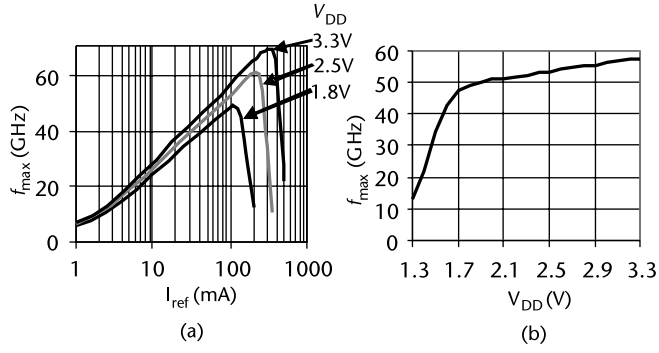

Figure 11.22 f_{max} for cascoded 3.3-V transistor: (a) versus I_{ref} and (b) versus V_{DD}.

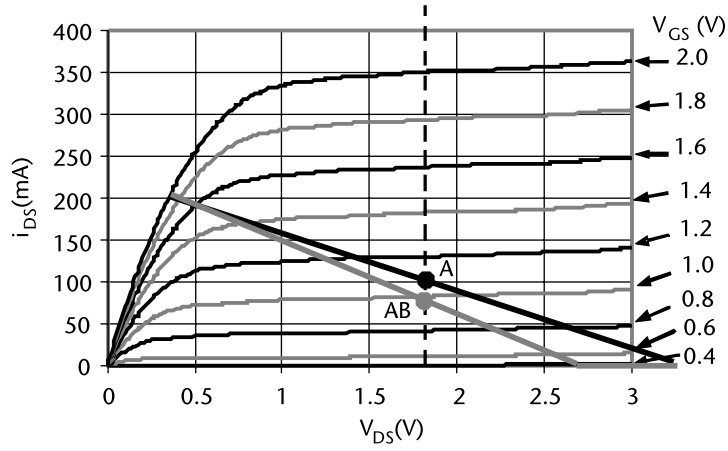

Figure 11.23 Transistor characteristic curves.

1.8 + 1.5 or 3.3V. A class AB amplifier would be biased at somewhat lower current, for example, at 75 mA. It can be assumed that the I_{DS} and V_{DS} swing is the same as for the class A amplifier. Note that if V_{GS} does not swing up to 2V as was assumed here, the V_{DS} swing is also affected. The required load resistance can be estimated as

$$P_o = \frac{v_{\text{rms}}^2}{R} \qquad R = \frac{v_{\text{rms}}^2}{P_o} = \frac{1.5^2}{2 \quad 0.075} = 15$$

A matching circuit will be required to convert 15 to 50 .

The current is determined from

$$P_o = \frac{v_p i_p}{2} \qquad i_p = \frac{2P_o}{v_p} = 0.1 \text{ or } 100 \text{ mA}$$

Thus, the transistors seem to be scaled appropriately to deliver 75 mW of power.

Now, using a peak transistor current of 200 mA, we can write

$$i_{C,peak} = I_{CC}(1 \quad \cos\theta) = 200 \text{ mA}$$

Using the nominal current of 75 mA,

$$I_{CC}\cos\theta = \quad 75 \text{ mA}$$

Then, solving for I_{CC},

$$I_{CC} = 125 \text{ mA}$$

We can find θ

$$\theta = \cos^{\quad 1}\left(\frac{75}{125}\right) = 127°$$

This allows us to calculate the efficiency

$$\eta = \frac{v_{op}}{V_{DD}} \frac{2\theta \quad \sin 2\theta}{4(\sin\theta \quad \theta\cos\theta)} = \frac{1.5}{1.8} \quad 0.632 = 0.53$$

For a CMOS amplifier, the gain may be as low as 3 dB, in which case we can predict the PAE as

$$PAE = \eta \ \ 1 \quad \frac{1}{G}\div = 0.265$$

Then with the basic design completed, we do some simulations. First, a quick small-signal simulation was done, primarily to check stability. Only minor stabilization is found to be necessary, primarily because the cascode structure reduces feedback and because of the low available gain with these 3.3-V devices. It was possible to make the amplifier unconditionally stable, either by reducing the input bias resistor, or by adding a small series output resistance. It was also noted that with small-signal matching and with a small-signal input, it was possible to achieve 15 dB of gain. However, with small-signal matching, the maximum achievable power added efficiency was less than 10% at an input power of 10 dBm and the output power was only about 14 dBm at this point. Thus, compared to the small-signal gain of 15 dB, the large signal gain has been reduced to about 4 dB.

The amplifier after stabilization (with the output resistor) and all approximate component values is shown in Figure 11.24. How these component values were determined will now be described.

Large signals simulations were performed in a series of iterations:

Step 1: Conjugately match the input, initially at the expected input power.
Step 2: Perform a load pull using the tuner at the previously determined input power level to determine the optimum output impedance. During the initial

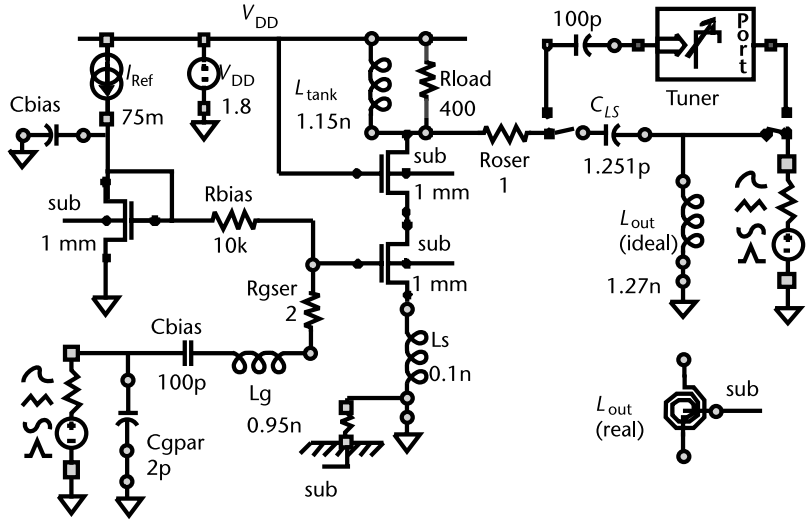

Figure 11.24 Amplifier schematic as used in simulation showing options to connect output components or a tuner to do load pulls.

load pull, the load inductance L_{tank} can be adjusted until the optimum load is close to the negative real axis on the Smith chart.

Step 3: Fix the output impedance at the optimal value and match the input again. In spite of enhanced reverse isolation with the cascode structure, the input matching may require minor adjustments whenever the output is changed.

Step 4: Sweep the input power to determine the optimum power level to achieve the highest efficiency.

Step 5: Repeat the steps several times until no further improvements in efficiency are seen.

For this example, the final optimal efficiency was about the same for input power levels of 12.5 dBm or 15 dBm. A load pull for an input power level of 12.5 dBm can be seen in Figure 11.25 showing the optimum impedance Z_{opt} to be $0.428 + j0$ normalized to 50 , or about 21.4 at an output power of 78.55 mW.

It can be noted that by selecting a large number of contours, the first contour shows only a slight reduction of power, in this case, down to 76.41 mW. This shows that although the optimum load is 21.4 , for any load impedance between about 18 and 27 , the power will still be over 76 mW meeting the initial requirement.

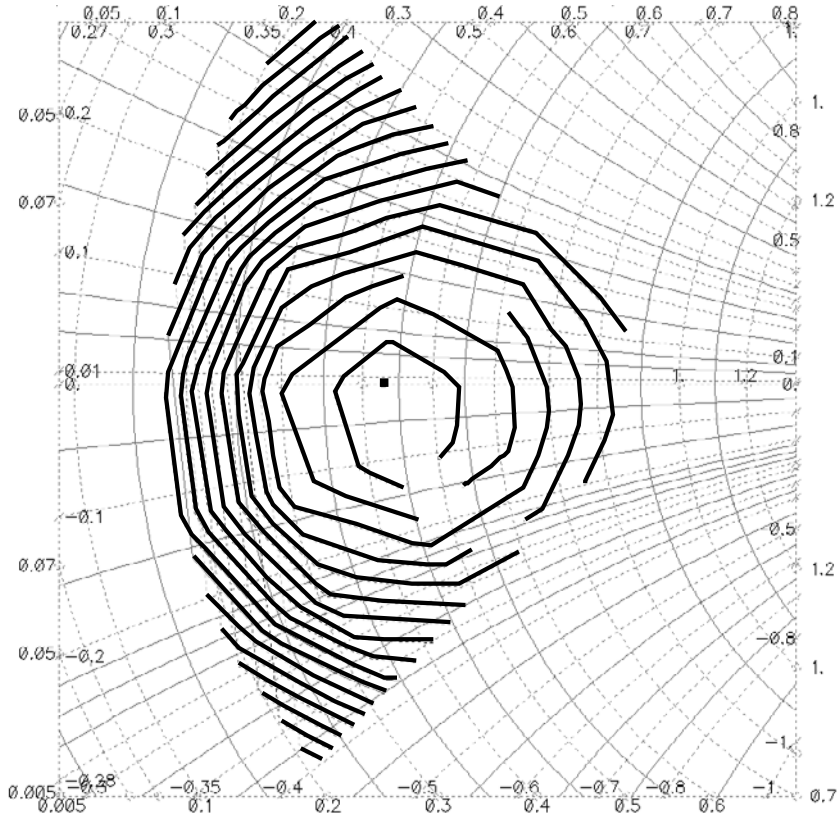

Figure 11.25 Load pull with 20 contours with phase sweep from 135° to 225° in steps of 15° and with magnitude sweep from 0.2 to 0.7 in five steps.

Similarly, with the input at 15 dBm (and conjugately matched), the load pull produced contours that were very similar except that the maximum output power was increased to 94.22 mW at nearly the same location on the Smith chart (about 21.4) with the first contour at 91.66 mW.

The tuner was then replaced with an LC network that would convert the 50 load to the same impedance of 21 by doing an S-parameter sweep of L_{out} and C_{LS} resulting in values of 1.27 nH and 1.251 pF. Then, L_{out} was replaced with a real spiral inductor with 2.5 turns, an outer diameter of 250 μm, and a track width of 25 μm, with a simulated inductance of 1.25 nH and a Q of 22.5. It was noted that the self-resonance frequency was at about 30 GHz, so this structure remains inductive to at least the fifth harmonic. Being an on-chip inductor, metal migration limits must be respected. Since the inductors are made out of thick top metal and 25-μm width is being used, and since there is no dc in this inductor, the current density was significantly less than the metal migration limit. In fact, with this relatively low power of 75 mW in this process with thick top metal lines, it turned out to be possible to realize all the inductors on chip.

Having completed the design including on chip inductors and optimally matched impedances, stability was checked again and it was seen that stability had improved (previously the minimum K_f was about 2; now it was 4). As a result, it would be possible to reduce or remove the stabilizing components with noticeable improvements in efficiency. Then, final large-signal simulations were performed. Figure 11.26 shows P_{out} and PAE versus P_{in} at 5.1, 5.2, and 5.3 GHz. The curves are all nearly superimposed showing that all three frequencies are about the same. The maximum efficiency is about 19% for inputs at 12.5 to 15 dBm for which the output power is 18 to 19 dBm. The PAE is low because of the transistor not being able to go to 0V and because the gain at the optimal PAE point is only about 3 to 4.5 dB.

Time-domain plots with input powers of 12.5 dBm and 15 dBm can be seen in Figures 11.27 and 11.28, showing voltage and current, respectively. The V_{D2} swing for a 12.5-dBm input is from 0.4V to 3.3V, while for a 15-dBm input, it is from 0.3V to 3.5V. In spite of being a cascoded structure, the voltage minimum is still

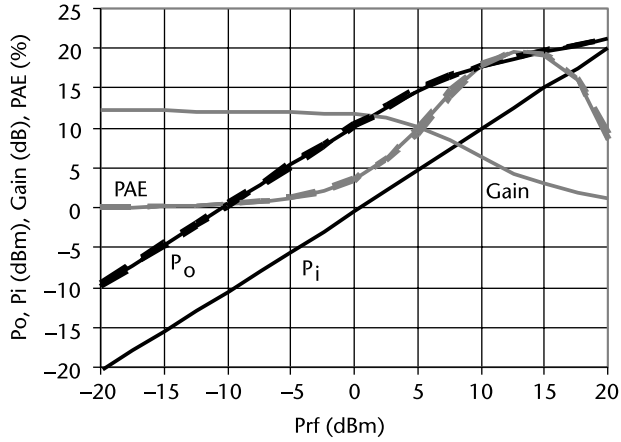

Figure 11.26 P_o versus P_i, PAE, and gain at three frequencies. The solid curve is at 5.2 GHz, and the dashed curves are at 5.1 and 5.3 GHz.

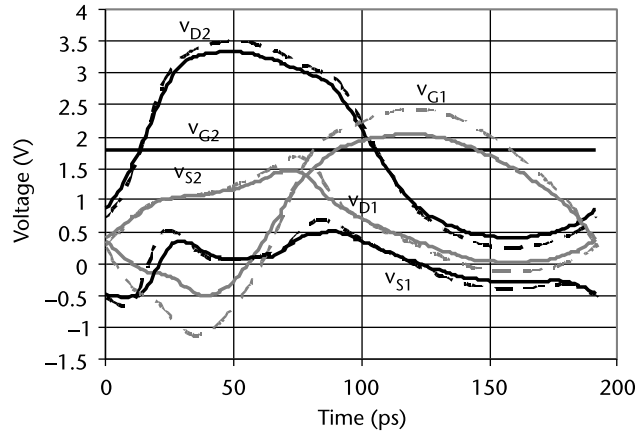

Figure 11.27 Time-domain curves for voltage. Solid curves are for P_{rf} of 12.5 dBm; dashed curves are for P_{rf} of 15 dBm.

very close to that predicted by the single transistor curve. This is possible because the presence of the source inductance allows the source voltage of the driver transistor to go negative. The current swing for a 12.5-dBm input is from 0 to 200 mA, or for a 15-dBm input it is from 0 to 215 mA. The estimate power using V_{pp} $I_{pp}/8$ results in 72.5 mW and 88.6 mW (or 18.6 and 19.5 dBm) for the two cases. An estimate for θ by assuming that the transistor is off during the time the current is negative results in 150° and 147°. Because θ is higher than predicted, the PAE is reduced from the earlier predicted value.

11.6 Class D Amplifiers

An example of a class D amplifier is shown in Figure 11.29(a) in which the switch alternately connects the input to ground or to V_{CC}. The output filter, consisting of L_o and C_o, is tuned to the fundamental frequency. This serves to remove the dc

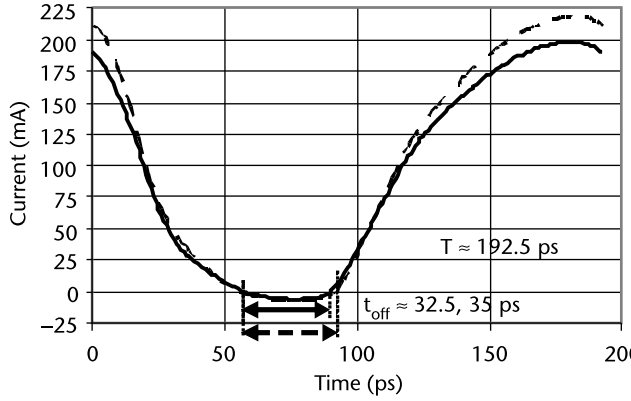

Figure 11.28 Time-domain curves for current. The solid curve is for P_{rf} of 12.5 dBm; the dashed curve is for P_{rf} of 15 dBm. Also shown is the approximate time that the transistors are off.

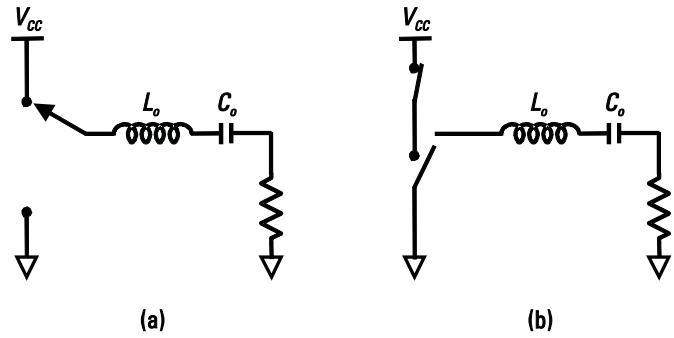

Figure 11.29 Class D amplifiers: (a) shown conceptually with a single switch and (b) as implemented with two switches driven with opposite phases.

component and the harmonics, resulting in a sine wave at the output. Figure 11.29(b) shows the class D amplifier as usually implemented with two separate switches. While the switch making the connection to ground is straightforward to implement, the switch connecting to V_{CC} is difficult to implement at RF. At lower frequencies, push-pull structures, often implemented with transformers, can be used to achieve a very high efficiency in the 10-MHz frequency range. In CMOS technologies, complementary transistors can be used for the two switches, but since PMOS transistors are considerably slower than NMOS transistors, class D amplifiers are rarely used in the gigahertz range. In the next section we will look at another switching amplifier, the class E amplifier, which is designed to operate with a switch to ground only. The class E amplifier has the high efficiency of the class D amplifier, but its simpler switching structure makes it feasible at high frequencies. For this reason, the class D amplifier will not be discussed further. For the interested reader, more information on class D amplifiers can be found in [5, 6].

11.7 Class E Amplifiers

The class E amplifier is shown with a bipolar transistor in Figure 11.30. It is designed to require a capacitor across the output of the transistor, which means that the capacitor C is the combination of the parasitic transistor output capacitor c_o and an actual added capacitor C_A. Thus, it is possible to obtain close to 100% efficiency even in the presence of parasitics.

11.7.1 Analysis of Class E Amplifier

Several simplifying assumptions are typically made in the analysis [4]:

1. The radio frequency choke (RFC) is large with the result that only dc current I_{dc} flows through it.
2. The Q of the output circuit consisting of L_o and C_o is high enough so that the output current i_o and output voltage v_o consist of only the fundamental component. That is, all harmonics are removed by this filter.

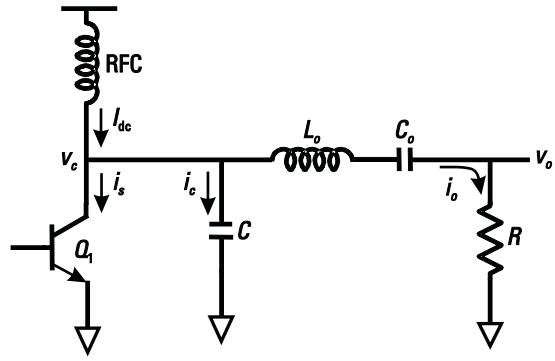

Figure 11.30 Class E amplifiers.

3. The transistor Q_1 behaves as perfect switch. When it is on, the collector voltage is 0V, and when it is off, the collector current is zero.
4. The transistor output capacitance c_o, and hence C, is independent of voltage

With the above approximations, the circuit can now be analyzed. Waveforms are shown in Figure 11.31. When the switch is on, the collector voltage is zero, and therefore the current i_C through the capacitor C is zero. In this case, the switch current $i_s = I_{\text{dc}} \quad i_o$. When the switch is off, $i_s = 0$. In this case, $i_C = I_{\text{dc}} \quad i_o$. This produces an increase of collector voltage v_C due to the charging of C. Due to resonance, this voltage will rise and then decrease again. To complete the cycle, as the switch turns on again, C is discharged and collector voltage goes back to zero again. If the component values are selected correctly, then the collector voltage will reach zero just at the instant the switch is closed, and as a result, there is no power dissipated in the transistor.

We cannot explicitly solve for voltage and current waveforms over the entire cycle. We can, however, determine the collector voltage waveform $v_c(\theta)$, where $\theta = \omega t$, by summing the currents at the transistor output. The voltage on this node will be related to the integral of the capacitor current and this can be solved [4] at a particular θ as:

$$v_C(\theta) = \frac{I_{\text{dc}}}{B} \quad y \quad \frac{\pi}{2} \div + \frac{V_{\text{om}}}{BR} \sin(\phi \quad y) + \frac{I_{\text{dc}}}{B}\theta + \frac{V_{\text{om}}}{BR} \cos(\theta + \phi) \tag{11.21}$$

where I_{dc} is the dc input current, I_{om} is the magnitude of the output current i_o, V_{om} is the magnitude of the output voltage and is given by the product of I_{om} and R, ϕ is the phase of v_o measured from the time the switch opens, $2y$ is the switch-off time in radians (e.g., $y = \pi/2$ for 50% duty cycle), and B is the admittance of the capacitance C.

The fundamental frequency component of $v_C(\theta)$ is $v_1(\theta)$. This is applied to $R + jX$ to determine output current, voltage, and power. Here jX is the residual impedance of the series combination of L_o and C_o, which are tuned to be slightly away from resonance at f_0.

For lossless components (as in the assumptions), the only loss is due to the discharge of C when the switch turns on. If the components are selected so v_C just

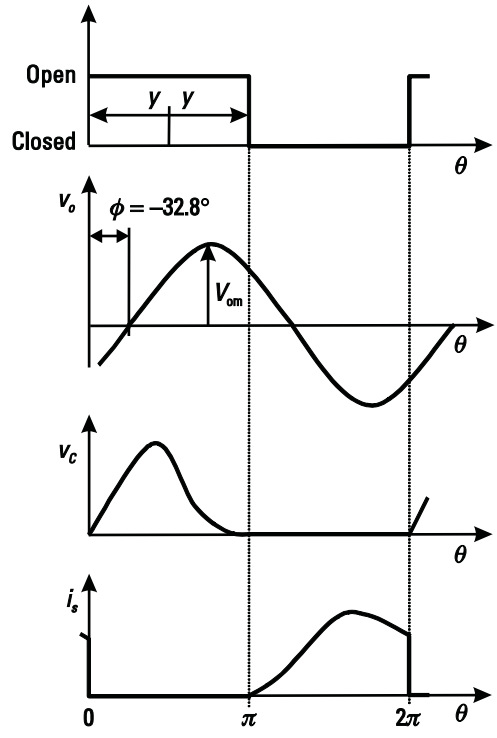

Figure 11.31 Class E waveforms.

reaches 0 as the switch turns on, no energy is lost and the efficiency is 100%. In practice, because the assumptions do not strictly hold and because components will not be ideal, the voltage will not be at zero and so energy will be lost. However, with careful design, efficiencies in the 80% range are feasible.

11.7.2 Class E Equations

It is necessary to choose B and X for the correct resonance to make sure $v_C = 0$ as the switch turns on and to make sure that $dv_C(\theta)/d\theta = 0$ to ensure that there is no current flowing into the capacitor.

Setting $v_C(\theta)$ and $d/d\theta v_C(\theta)$ to 0 at $\theta = \pi/2 + y$ results in

$$\begin{aligned} \phi &= 32.48° \\ B &= \frac{0.1836}{R} \\ X &= 1.152R \end{aligned} \tag{11.22}$$

It can be shown that

$$v_{om} = \frac{2}{\sqrt{1+\pi^2/4}} V_{CC} \quad 1.074V_{CC} \tag{11.23}$$

and that

$$P_o \frac{2}{1+\pi^2/4}\frac{V_{CC}^2}{R} \quad 0.577\frac{V_{CC}^2}{R} \tag{11.24}$$

The dc current is given by

$$I_{dc} = \frac{V_{CC}}{1.734R} \tag{11.25}$$

The peak transistor currents and voltages are given by

$$\begin{aligned} v_{C,peak} &= 3.56\, V_{CC} \\ i_{s,peak} &= 2.86 I_{dc} \end{aligned} \tag{11.26}$$

The resulting output power is 78% of the power produced by a class B PA, but the efficiency approaches 100%.

11.7.3 Class E Equations for Finite Output *Q*

If the Q of the output circuit is not infinity as initially assumed, but more typically less than 10, then some harmonic current will flow. This can result in the collector voltage not being zero at the instant the switch closes. Formulas for optimum operation in this case were shown by Sokal [8]:

$$X = \frac{1.110Q}{Q \quad 0.67}R \tag{11.27}$$

$$B = \frac{0.1836}{R} \quad 1+\frac{0.81Q}{Q^2+4}\div \tag{11.28}$$

$$Q \quad \frac{\omega_0 L}{R} \tag{11.29}$$

Then we may need to insert a filter between C_o and R to prevent excessive harmonic currents from reaching R.

11.7.4 Saturation Voltage and Resistance

Previously it was assumed that the output voltage would be 0 when the transistor was on. In reality, it will be equal to the transistor saturation voltage V_{SAT}. As described in [4], this can be accounted for by replacing the power supply voltage with $V_{eff} = V_{CC} - V_{SAT}$ for all calculations, except power input. As for the transistor having nonzero on resistance R_{on}, this can be accounted for by changing the value of V_{eff} by $V_{eff} \quad R/(R + 1.365R_{on})V_{CC}$. This is valid for 50% duty cycle.

11.7.5 Transition Time

Ideally, no power is dissipated during the transition between off and on. The turn-on transition for nonoptimum conditions can be approximated with a linear ramp of current. This produces a parabolic collector voltage waveform. As described in [4], the current and voltage waveforms can be integrated to determine dissipated power P_{dT}.

$$P_{dT} = \frac{1}{12}\theta_S^2 P_o \tag{11.30}$$

where θ_S is the transition time in radians and P_o is the output power. Then efficiency is given by

$$\eta = 1 \quad \frac{1}{12}\theta_S \tag{11.31}$$

All the above losses due to saturation voltage, on resistance and turn-on transient, can be combined by summing dissipated power or by finding each efficiency by itself and then multiplying the efficiencies.

Further information and detailed examples of class E amplifier designs are shown by Cripps [5] and by Albulet [6].

Example 11.4: Class E Amplifier

Design a class E amplifier that delivers 200 mW from a 3-V power supply at 2.4 GHz. Assume an ideal transistor and aim for a Q of 3 for the output circuit. Specify device ratings and components.

Solution:

Using $P_o \quad 0.577V_{CC}^2/R$ results in $R = 26 \quad$. The maximum transistor voltage is $v_{C,\max} = 3.56 \quad 3 = 10.68$V. If this large voltage is not permissible (and it is quite likely that it is not), the power supply voltage may need to be reduced.

$$I_{dc} = \frac{V_{CC}}{1.734R} = \frac{3}{1.734 \quad 26} = 66.6 \text{ mA}$$

$$i_{c,\text{peak}} = 2.86I_{dc} = 190.5 \text{ mA}$$

From (11.28),

$$B = \frac{0.1836}{R} \quad 1 + \frac{0.81Q}{Q^2 + 4} \div = \frac{0.1836}{26} \quad 1 + \frac{0.81 \times 3}{3^2 + 4} \div = 0.00838$$

Hence $C = 0.50$ pF. Since $Q = 3$, C_o has a reactance of 78 and is therefore 0.85 pF. Using (11.27),

$$X = \frac{1.110Q}{Q \quad 0.67}R = \frac{1.110 \times 3}{3 \quad 0.67}26 = 37.2$$

L_o therefore has a reactance of 37.2 + 78 = 115.2 and is thus 7.64 nH. Ideally, the RFC should have a reactance of at least 10R and thus should be at least 17 nH, which would likely need to be an off-chip inductor.

This circuit was simulated using a bipolar process with f_T that is 25 times higher than the operating frequency. With numbers calculated as above, and choosing a transistor size that has optimal f_T at about 66 mA, the results in Figure 11.32 were obtained. For simplicity, the input was a ±1.5-V pulse waveform through a 50 source resistance. In a real circuit, a more realistic input waveform would have to be used.

It can be seen that the output transistor collector voltage is still decreasing and has not reached its final value close to zero when the transistor switches on. The problem is the parasitic output capacitance of the very large transistor. As a result, the output power is only about 77 mW and dc power is of the order of 100 mW. As a first-order compensation, the capacitor C can be reduced to compensate. With this done, the results are as shown in Figure 11.33.

With this adjustment, the results are now close to the predicted values. The average current is about 60 mA; collector output voltage is just over 12V, a little bit more than predicted; the collector current peaks at 180 mA, close to the predicted value; and the output voltage is about 5.8V peak to peak, nearly the predicted value. The output power is 162 mW, while the dc power is about 180 mW for a

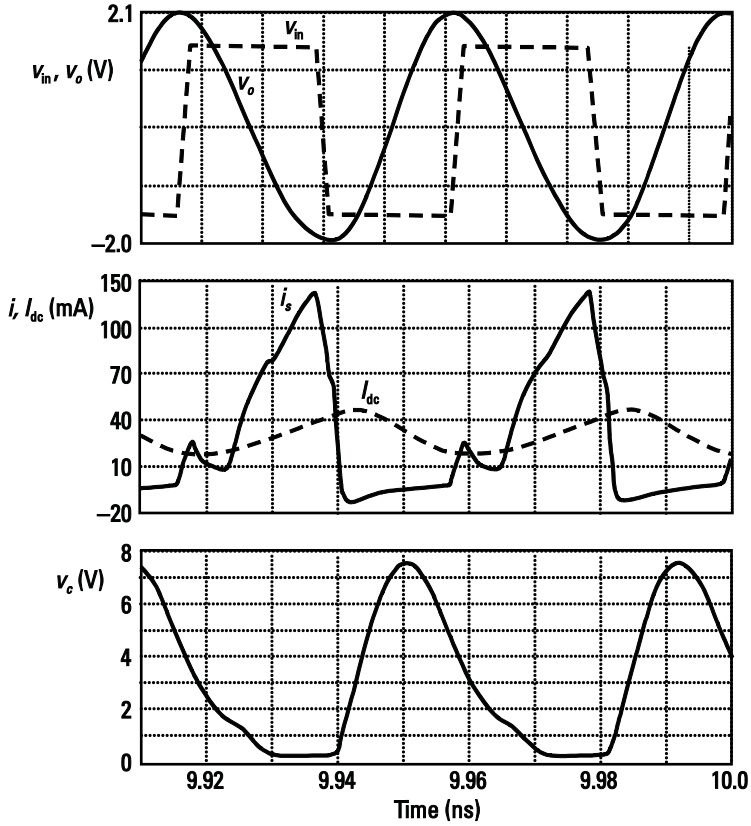

Figure 11.32 Initial simulated class E waveforms.

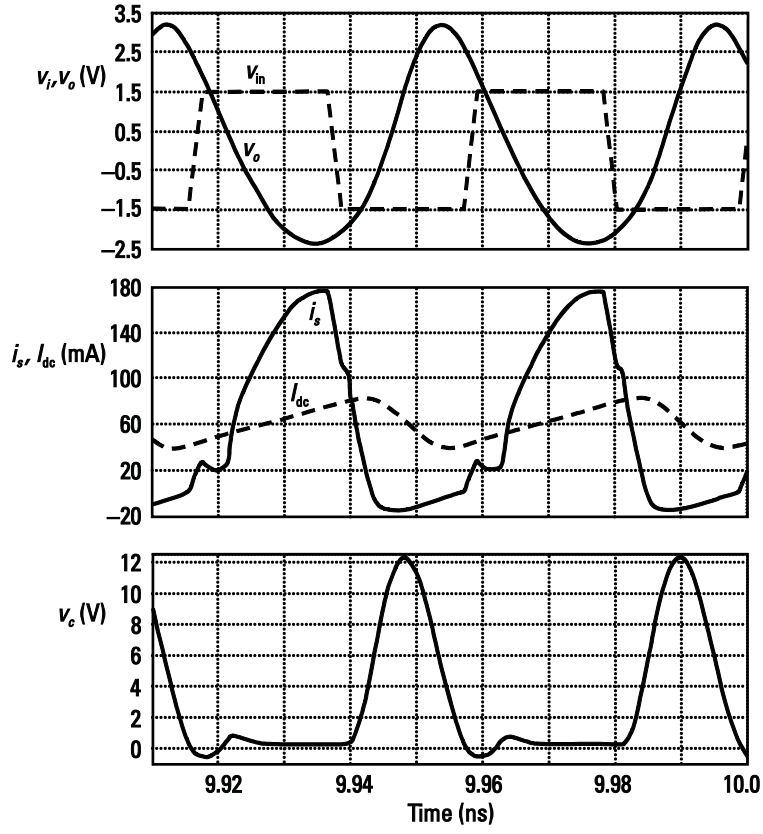

Figure 11.33 Simulated class E waveforms with reduced capacitance *C*.

dc-to-RF efficiency of about 90%. However, with the unrealistic pulse input, a significant amount of power is fed into the input so PAE will be lower; in this example, with an input current of nearly 20 mA, PAE is estimated to be about 75%.

11.8 Class F Amplifiers

In the class F amplifier shown with a bipolar transistor in Figure 11.34, an additional resonator is used, with the result that an additional harmonic, typically the third harmonic, is added to the fundamental in order to produce a collector voltage more like a square wave. This means the collector voltage is lower while current is flowing, but higher while the current is not flowing, so the overall efficiency is higher.

The typical waveforms for a class F amplifier are shown in Figure 11.35, where the collector voltage has a slightly flattened appearance while the output voltage is sinusoidal. Current only flows for half the time or less in order to ensure that there are third-harmonic components and to maximize the efficiency (zero current while there is a finite collector voltage).

The transistor behaves as a current source producing a half sinusoid of current $i_C(\theta)$, similar to class B operation. L_o and C_o make sure the output is a sinusoid. The third-harmonic resonator (L_3, C_3), causes a third-harmonic component in the

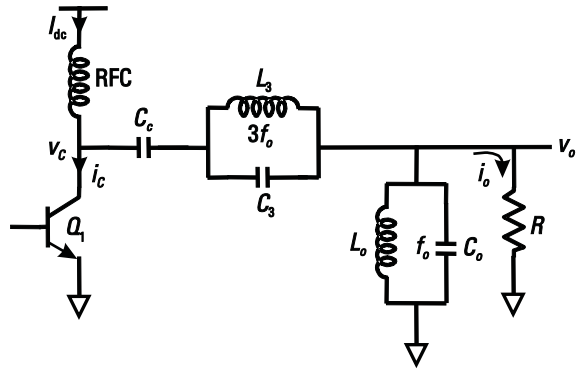

Figure 11.34 Class F amplifier.

collector voltage. At the correct amplitude and phase, this third-harmonic component produces a flattening of v_C as shown in Figure 11.36. This results in higher efficiency and higher output power.

If the amplitude of the fundamental component of the collector voltage is V_{cm} and the amplitude of the third harmonic is V_{cm3}, then it can be shown that maximum flatness is obtained when $V_{\text{cm3}} = V_{\text{cm}}/9$.

Thus, with

$$V_{\text{cm3}} = \frac{V_{\text{cm}}}{9} \tag{11.32}$$

it can be seen from Figure 11.36 that the peak collector voltage is

$$\frac{8}{9}V_{\text{cm}} = V_{\text{CC}} \quad V_{\text{cm}} = \frac{8}{9}V_{\text{CC}} \tag{11.33}$$

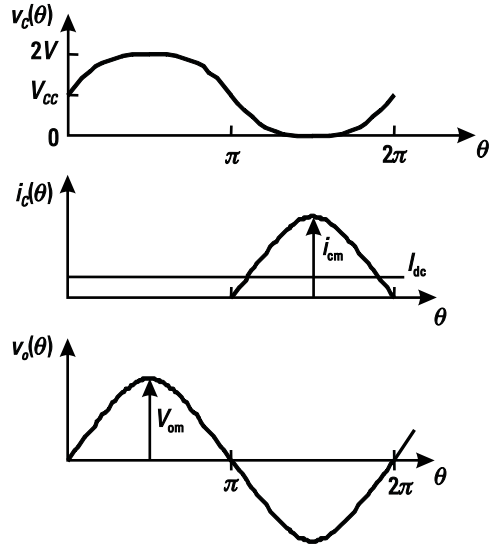

Figure 11.35 Class F waveforms.

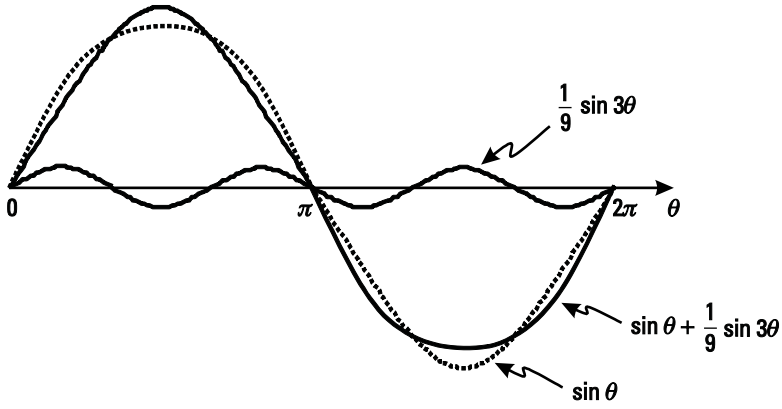

Figure 11.36 Class F frequency components of waveforms.

As an aside comment, the Fourier series for the ideal square wave is

$$\sin\theta + \frac{1}{3}\sin\theta + \frac{1}{5}\sin 5\theta \ldots \tag{11.34}$$

However, choosing $V_{cm3} = 1/3\,V_{cm}$ would produce a nonflat waveform, as shown in Figure 11.37.

The efficiency can be calculated as P_o/P_{dc}. By taking a Fourier series of the $i_C(\theta)$ waveform with a peak value of i_{cm}, it can be shown that the dc value I_{dc} is equal to i_{cm}/π and the fundamental value of the output current i_o is equal to $i_{cm}/2$. As well, V_{om} is equal to V_{cm}. Thus, efficiency can be calculated as

$$\frac{P_o}{P_{dc}} = \frac{\frac{1}{2}(i_o \times V_{om})}{I_{dc}V_{CC}} = \frac{\frac{1}{2}\;\frac{i_{cm}}{2}\times\frac{9}{8}V_{CC}\div}{\frac{i_{cm}}{\pi}\times V_{CC}} = \frac{9}{8}\times\frac{\pi}{4} \qquad 88.4\% \tag{11.35}$$

11.8.1 Variation on Class F: Second-Harmonic Peaking

A second resonator allows the introduction of a second-harmonic voltage into the collector voltage waveform, producing an approximation of a half sinusoid, as seen in Figure 11.38. It can be shown that the amplitude of the second-harmonic voltage should be a quarter of the fundamental voltage.

It can be shown that the peak output voltage is given by

Figure 11.37 Fundamental and third harmonic to make a square wave.

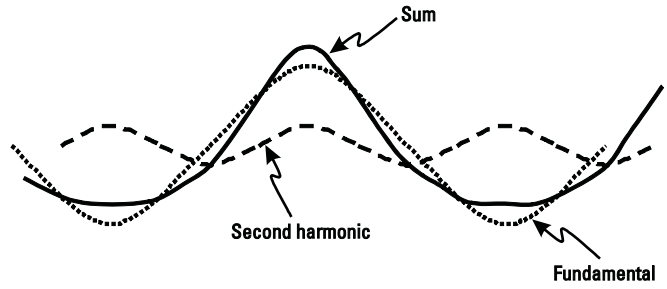

Figure 11.38 Second-harmonic peaking waveforms.

$$V_{\text{om}} = \frac{4}{3} V_{\text{CC}} \tag{11.36}$$

and the efficiency is given by

$$\eta = \frac{8}{3}\pi \quad 84.9\% \tag{11.37}$$

11.8.2 Variation on Class F: Quarter-Wave Transmission Line

A class F amplifier can also be built with a quarter-wave transmission line as shown in Figure 11.39, with waveforms shown in Figure 11.40.

A quarter-wavelength transmission line transforms an open circuit into a short circuit and a short circuit into an open circuit. At the center frequency, the tuned circuit (L_o and C_o) is an open circuit, but at all other frequencies, the impedance is close to zero. Thus, at the fundamental frequency the impedance into the transmission line is R_L. At even harmonics, the quarter-wave transmission line leaves the short circuit as a short circuit. At odd harmonics, the short circuit is transformed into an open circuit. This is equivalent to having a resonator at all odd harmonics,

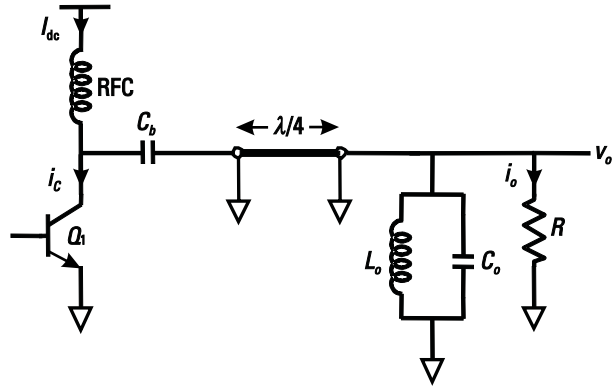

Figure 11.39 Transmission line in a class F amplifier.

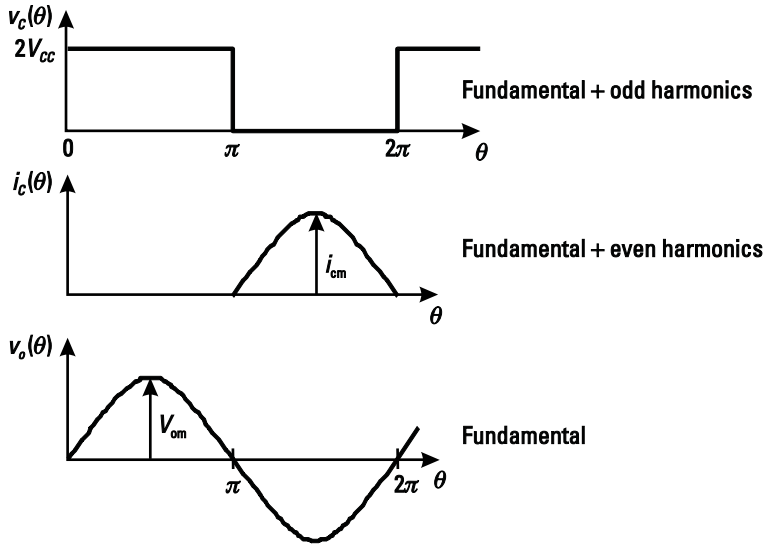

Figure 11.40 Waveforms for a class F amplifier with a transmission line.

with the result that the collector voltage waveform is a square wave (assuming that the odd harmonics are at the right levels).

The collector current consists of the fundamental component (due to the load resistor) and all even harmonics. We note there are no odd harmonics, since current cannot flow into an open circuit. This produces a half sinusoid of current.

Only the fundamental has both voltage and current; thus, power is generated only at the fundamental. As a result, this circuit ideally has an efficiency of 100%.

Saturation voltage and on resistance can be accounted for in the same way as for the class E amplifier.

Example 11.5: Class F Power Amplifier

Design a class F amplifier with third-harmonic peaking to deliver 200 mW from a 3-V supply.

Solution:

The maximum output voltage can be determined.

$$V_{\mathrm{CC}} = \frac{8}{9} V_{\mathrm{om}} \qquad V_{\mathrm{om}} = \frac{9}{8} V_{\mathrm{CC}} = 3.375\mathrm{V}$$

The required output resistance is found as

$$\frac{V_{\mathrm{om}}^2}{2R} = 0.2 \qquad R = \frac{3.375^2}{2 \quad 0.2} = 28.5$$

The maximum collector voltage swing is $V_{\max} = 2V_{CC} = 6\mathrm{V}$.

The peak output current is $i_{o,\mathrm{peak}} = V_{\mathrm{om}}/R = 3.375/28.5 = 118.4$ mA

$$i_{\mathrm{cm}} = 2 \quad i_{o,\mathrm{peak}} = 237 \text{ mA}$$

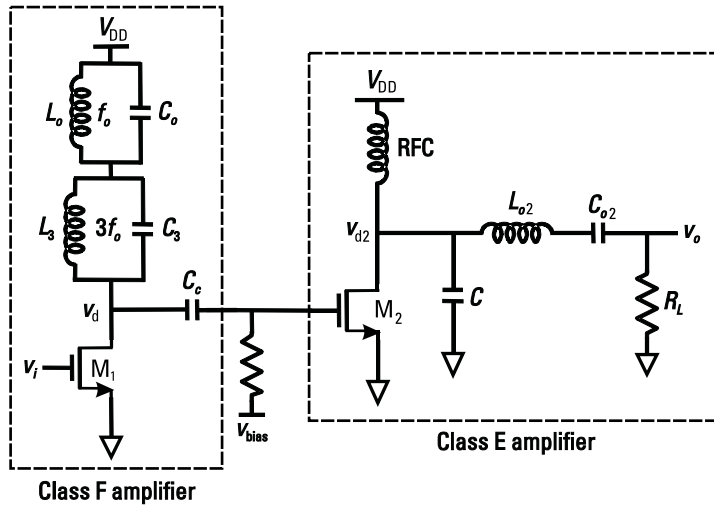

Figure 11.41 Class F amplifier driving a class E amplifier

$$I_{dc} = i_{cm}/\pi = 0.237/\pi = 75.5 \text{ mA}$$

Check powers, efficiencies $P_{dc} = I_{dc}\, V_{CC} = 0.0755 \quad 3 = 0.2264\text{W}$

$$\eta = \frac{P_o}{P_{dc}} = \frac{0.20}{0.2264} \qquad 88.4\%$$

Of course, in a real implementation, efficiency would be lower because of losses due to saturation voltage, on resistance, finite inductor Q, imperfect RFC, and parasitics.

One useful application of a class F amplifier is as a driver for the class E amplifier, as shown in Figure 11.41. A class E amplifier is ideally driven by a square wave. Such a waveform is conveniently available on the collector of the class F amplifier, so this node is used to drive the class E amplifier. Note that although shown with CMOS transistors, bipolar can also be used.

11.9 Class G and H Amplifiers

The class G amplifier shown in Figure 11.42 has been used mainly for audio applications, although recently variations of this structure have been used at megahertz frequencies for signals with high peak-to-average ratios (high crest factor), for example, in digital telephony applications.

This topology uses amplifiers powered from different supplies. For low-level signals, the lower supply is used and the other amplifier is disabled.

Class H uses a linear amplifier, such as a push-pull class B amplifier as shown in Figure 11.43, to amplify the signal. However, its power supplies track the input signal or the desired output signal. Thus, power dissipated is low, since the driver transistors are operated with a low-voltage V_{CE}. As a result, the efficiency can be much higher than for a class A amplifier.

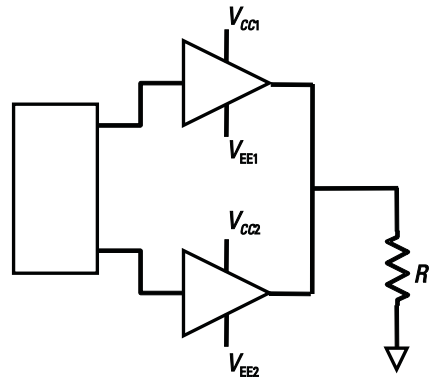

Figure 11.42 Class G amplifier.

The power supplies use a highly efficient switching amplifier. Noise (from switching) is minimized by the power supply rejection of the linear amplifier.

As with the class G amplifier, this technique has mainly been used for lower frequencies. However, this technique can be modified so that the power supply follows the envelope of the signal, rather than the signal itself. Discussion of such circuits for code division multiple access (CDMA) RF applications can be found in [2].

11.10 Summary of Amplifier Classes for RF Integrated Circuits

Classes D, G, and H are not appropriate for RF integrated circuits in the gigahertz range, so will not be included in this summary.

The main advantage that the class A amplifier has is its linearity, although good linearity can also be achieved with class AB if the power is backed off, thereby sacrificing efficiency for linearity. Thus, in spite of reduced efficiencies, these amplifiers are used for low power applications where efficiency is less important, or in applications requiring linearity, for example, in quadrature amplitude modulation (QAM),

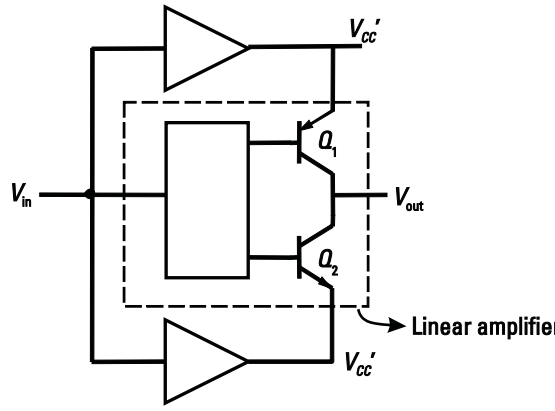

Figure 11.43 Class H amplifier.

where the amplitude is not constant. Linearization techniques, to be discussed later, are not yet widely used for fully integrated power amplifiers.

Class A, AB, and B amplifiers can achieve approximately the same fundamental RF output power. However, class B has a theoretical maximum efficiency of 78%, while class A has a theoretical maximum efficiency of 50%. It should be noted that practical efficiencies are much lower in fully integrated power amplifiers due to a number of nonidealities such as finite inductor Q, saturation voltage in the transistors, and tuning errors due to process, voltage, or temperature variations. Efficiencies of half of the theoretical maximum would be considered extremely good, especially in a low-voltage process.

Class C amplifiers have theoretical maximum efficiencies higher than the class B, approaching 100% as the conduction angle decreases. However, this increase in efficiency is accompanied by a decrease in the output power. The output power approaches zero as the efficiency approaches 100%. Because of the difficulty of achieving low conduction angles on an integrated circuit, and because of other losses, completely integrated class C amplifiers are very rarely seen.

The class F amplifier can be seen as an improvement over class C or single-ended class B amplifiers in terms of output power (theoretically up to 27% higher than class B) and efficiency (theoretically 88% versus 78% for class B). However, the added resonant circuits have loss due to finite Q and result in an increase of complexity and larger chip area.

Class E can operate at radio frequencies because the resonant circuits are tuned to help the transistor switch. While efficiencies up in the 90% range can be achieved for hybrid designs, fully integrated designs have additional losses due to low-quality passives, such as inductors, so efficiencies of 60% are considered quite good. Output power is typically a few decibels lower than for similarly designed class AB amplifiers. While class AB amplifiers might have a maximum transistor voltage of about twice the power supply voltage, class E can have a swing higher than three times the power supply voltage. The maximum supply voltage and the breakdown voltage have been reduced for each new generation of process, with a typical value now being 1.8V or less. Thus, for class E amplifiers, if the fastest transistors are used, the supply voltage must typically be set to less than the maximum supply voltage. Many CMOS processes have options allowing thicker oxide transistors, which have a higher breakdown voltage at the expense of speed (lower f_T, $f_{\max}$), and increasingly such transistors must be used for power amplifier design.

Other techniques exist for combining two amplifiers with different output power and different peak operating conditions and then optimizing the combination for improved performance compared to a single amplifier. One possible optimization allows high efficiency over a broader range of input power [5]. While these techniques show promise, it still remains for someone to exploit these for integrated power amplifiers.

11.11 AC Load Line

The ac or dynamic load line shows the excursion of current versus voltage at the operating frequency. As shown in Figure 11.44, because of reactive impedance,

voltage and current will be out of phase and the dynamic load line will no longer be a straight line, appearing instead as an ellipse.

Figure 11.44 shows a simulated family of curves for increasing input amplitude from a 25-mV to a 200-mV peak. The amplifier is shown with an inductive output impedance resulting in a current that lags voltage. In addition, for nonlinear circuits, with harmonics, the current versus voltage characteristics will no longer be a simple ellipse, and patterns that are more complex can be seen. An example is shown in Figure 11.45. In this example, the load is now tuned so that current and voltage are in phase. Inputs of 200 and 400 mV are applied, resulting in current and voltage, which are visibly nonlinear, resulting in dynamic load lines with loops in the characteristics.

11.12 Matching to Achieve Desired Power

Given a particular power supply voltage and resistance value, the achievable amount of power is limited by $P_o \quad V^2_{CC}/2R$. Obviously, R must be decreased to achieve higher P_o. This is achieved by an impedance transformation at the output, as illustrated in Figure 11.46. As an example, if $V_{CC} = 3$V and $P_o = 1$W, the required

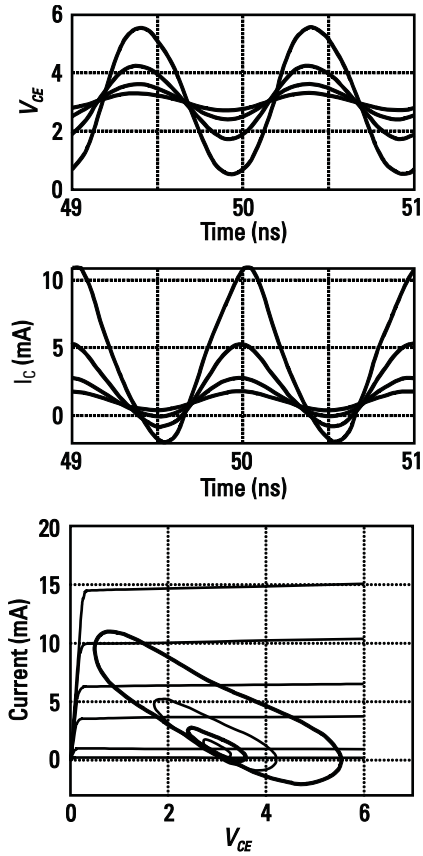

Figure 11.44 Time-domain waveforms and dynamic load line for reactive circuits.

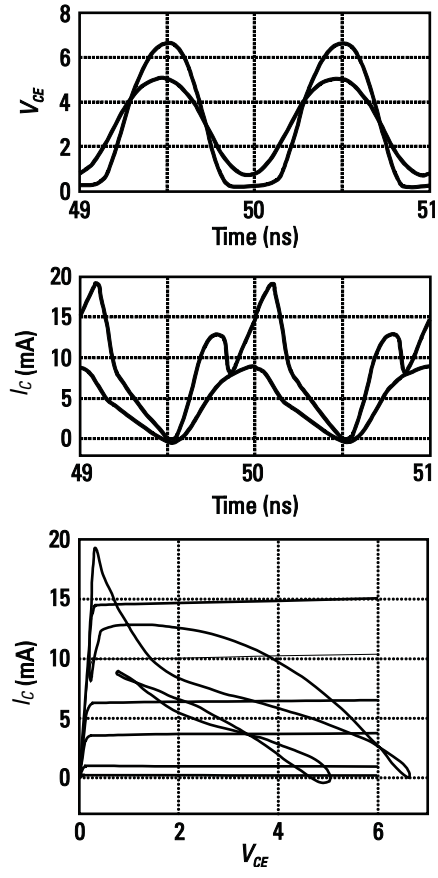

Figure 11.45 Time-domain waveforms and dynamic load line for large signals.

resistance R is approximately 4.5 . Similarly, for a P_o of 500 mW, R needs to be approximately 9 , and for a P_o of 200 mW, R needs to be about 22.5 .

Generally, the smaller R is, compared to R_L, the narrower will be the bandwidth of the circuit; that is, the amplifier will be able to produce useful output power over a narrower band of frequencies. It is possible to increase the bandwidth by using a higher order of matching network; for example, instead of an *ell* network, a *double ell* network can be used to convert first to an intermediate impedance, usually $R_{\text{int}} = \sqrt{RR_L}$, and then to the final value as shown in Figure 11.47. This choice of R_{int} maximizes the bandwidth. As an example, if $R = 4.5$ and $R_L = 50$, the optimal R_{int} is about 15 .

Sometimes, higher Q is desired, and then an intermediate resistance higher than R_L might be used. A possible matching network to achieve this is illustrated in Figure 11.48.

Bond wire inductance can be used for realizing series inductance. Examples of series inductance can be found in impedance transformation networks or in the output of the class E amplifier. Bond wire inductance has the advantage of high power handling capability and high Q compared to integrated inductors.

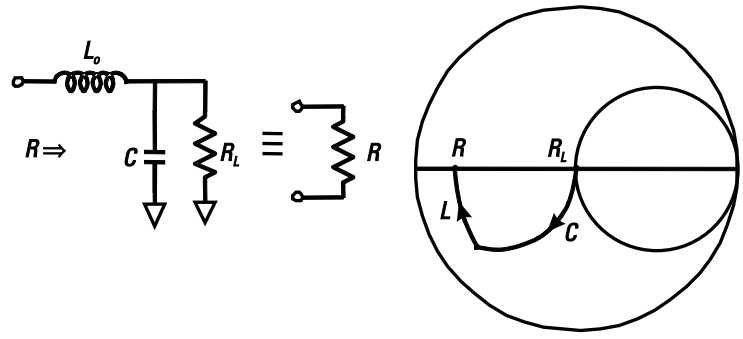

Figure 11.46 Matching example.

11.13 Transistor Saturation

Efficiency increases rapidly with increasing input signal level until saturation of the input device occurs. After the transistor saturation, efficiency is fairly constant, but drops somewhat due to gain compression and the resulting increase in input power. Bipolar transistors used in class A, B, and C amplifiers are usually operated just into saturation to maximize efficiency, while CMOS transistors are operated just into the triode region. Class E (and sometimes F) is operated as a switch, between saturation and cutoff. There may be difficulty in properly modeling the switching transistor, which increases the design difficulty. It can be noted that power series approximations are not particularly good for a transistor that is switching hard. An important design consideration for operation into saturation is that proper base drive is required to remove stored charge to get a transistor out of saturation fast.

11.14 Current Limits

As described earlier, with a 3-V power supply, for an output power of 500 mW, the load resistance must be about 9 and the required current is 333 mA. Because of efficiency issues and because the transistor is on for some reduced time, the peak

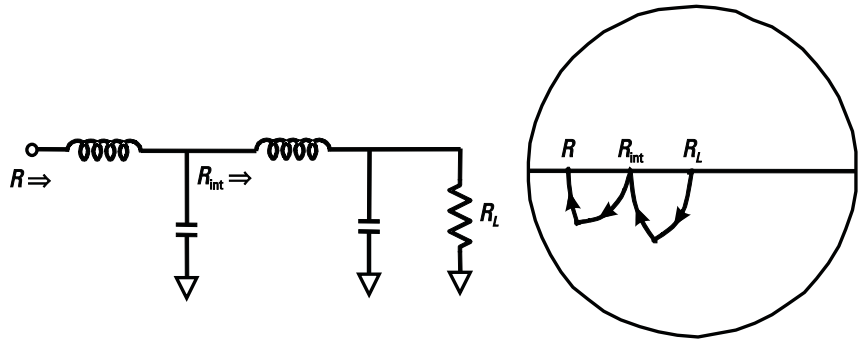

Figure 11.47 Broadband matching circuit.

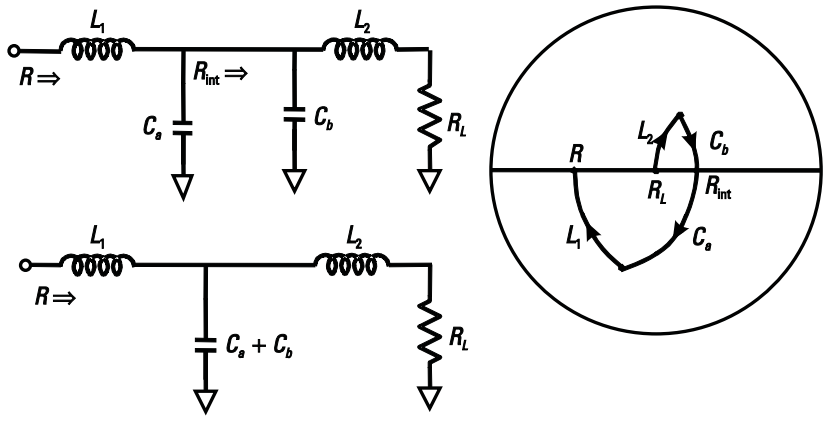

Figure 11.48 High-Q matching circuit.

collector current can easily be over 1A. As a result, there is a requirement for huge transistors with very high current handling requirements. This requires the use of transistors with multiple fingers as shown in Figure 11.49 for a bipolar transistor, as well as multiple transistors distributed to reduce the concentration of heat and to reduce the current density. However, the use of transistors with multiple fingers and multiple transistors introduces the new concern of making sure each finger and each transistor is treated the same as every other finger and transistor. This is important in order to make sure that the connection to and from each transistor is exactly the same to avoid mismatch of phase shifts and, in the case of bipolar transistors, to avoid local hot spots or thermal runaway. Power combining will be discussed further in Section 11.16, and thermal runaway is discussed further in Section 11.17.

Also, for such high currents, metal lines have to be made wide to avoid problems with metal migration, as described in Section 6.6.

As for transistors, a large transistor cannot become too long or it will not be able to handle its own current. With bipolar transistors, the current handling capability is directly proportional to the emitter area, while for CMOS transistors it is proportional to the transistor width. Thus, for a bipolar transistor as the emitter

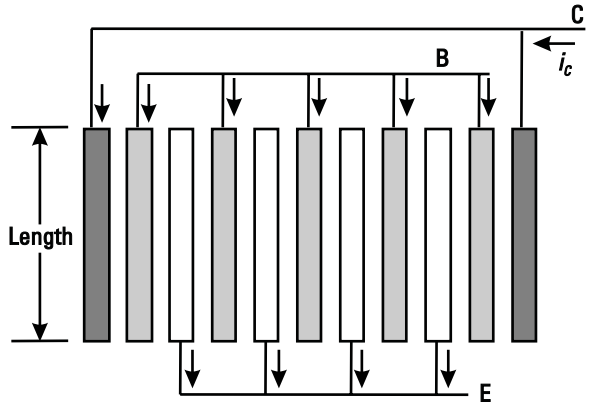

Figure 11.49 Bipolar transistor with multiple stripes.

length is doubled, or for a CMOS transistor as the transistor width is doubled, the current is doubled. However, if the line width stays the same, the maximum current capability is the same. As an example, if the transistor current is 1 mA and line widths are appropriate to accommodate this current, if the transistor size is doubled, the same line widths may not be sufficient. However, to increase current capability, it is possible to use multiple metal layers, for example, metals 1, 2, and 3. Higher metals are often thicker, resulting in a higher total current capability. In this example, the original maximum current of 1 mA might be increased to 2 mA or even 4 mA. Another point to keep in mind is that since current flows from collector to emitter in a bipolar transistor, or from drain to source in a CMOS transistor, the current density in the emitters or the sources is highest close to the external emitter or source contact, which for Figure 11.49 is on the bottom. A trade-off with stacking multiple layers of metal on emitters and collectors, or sources and drains, is that the capacitance is increased; hence, frequency response will be compromised.

11.15 Current Limits in Integrated Inductors

Integrated inductors as used for LNAs and oscillators are typically 5 to 25 μm wide. Typically, higher-frequency designs make use of narrower lines in order to decrease parasitic capacitance with resulting higher self-resonant frequencies. Also at higher frequencies, the required inductance is typically smaller; hence, the total inductor length is shorter, compensating for the higher resistance per unit length found in narrower lines. In a modern process, top metal may be very thick, some processes may make use of copper, and some inductors allow for two layers to be strapped together. Hence, in some processes the current handling capability has greatly improved, in some cases up to 10 mA/micron or more, allowing 250-mA dc or more for a 25-micron-wide inductor, with typically about four times more ac current allowed. In processes that only allow 1 to 2 mA/micron, the same line could handle no more than 25 to 50 mA of dc current. This would obviously limit the ability to do on-chip tuning or matching for power amplifiers.

Recently, slab inductors or slab transformers making use of wide strips of metal have been used, allowing up to 35-dBm output from a fully integrated CMOS power amplifier [9].

11.16 Power Combining

For high power, it is possible to combine multiple transistors at the output as shown in Figure 11.50. This distributes the heat and limits the current density in each transistor (compared to a single super-huge transistor).

However, with many transistors, the base drive to the outside transistors can be phase delayed compared to the shortest path, so it is important to keep the line lengths equal, as illustrated in Figure 11.50. However, this is a trade-off with the additional parasitic capacitance of the lines, so many designers choose not to employ this technique fully. Note also that as with all RF or microwave circuits, sharp bends are to be avoided. Line delay or phase shift can be determined by considering that

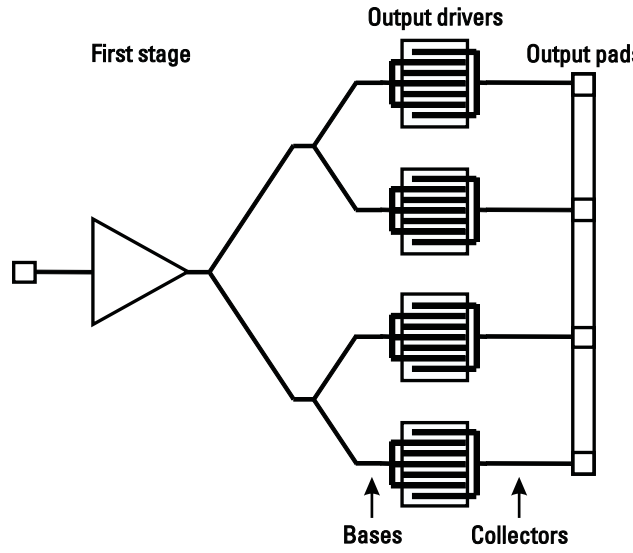

Figure 11.50 Multiple transistors.

the wavelength of a 1-GHz sine wave in free space is 30 cm. This results in a phase shift of about 1.2°/mm. Wavelength is inversely proportional to frequency and inversely proportional to the square root of the dielectric constant ε_R. Thus, for SiO_2 with ε_R of about 4 and at 5 GHz, we now have a phase shift of about 12°/mm. At 5 GHz, for a distance of 5 mm we have a phase shift of about 60°. This is obviously of critical importance, especially when considering that sometimes an exact phase shift is required, such as the 32° of phase shift required for a class E amplifier.

Note that it is possible to use multiple output pads for parallel bond wires, or instead to have a "wide" pad, shown in Figure 11.50, to connect more bond wires if desired. A typical requirement is that a width between 60 and 90 μm is required for each bond wire.

Instead of connecting all transistors together, then transforming the output impedance, in some recent PA designs, each transistor has been connected to an individual transformer or to a distributed transformer [9–11]. This provides output coupling and impedance matching while keeping the transistor output voltage low enough to avoid breakdown. The key to this working is the evolution of better transformer modeling tools.

Power combining can also be done off chip, for example, using various types of couplers such as the backward wave couplers (stripline overlay, microstrip Lange) for octave bandwidth, or the branch-line coupled amplifier. These will not be discussed further here, but information is found in many texts, for example, [12].

11.17 Thermal Runaway—Ballasting

Under high power, the temperature will increase. With constant base-emitter voltage, current increases with temperature. Equivalently, with constant current, base-emitter voltage decreases with temperature as

$$\left.\frac{V_{BE}}{T}\right|_{I=\text{constant}} \quad 2\frac{\text{mV}}{^{\circ}\text{C}} \tag{11.38}$$

Thus, if V_{BE} is held constant, if temperature increases, current increases, and as a result, more power is dissipated and temperature will increase even more. This phenomenon is known as *thermal runaway*. Furthermore, for unbalanced transistors, the transistor with the highest current will tend to be the warmest and hence will take an even higher proportion of the current. As a result, it is possible that the circuit will fail. Typically, ballast resistors are added in the emitters as a feedback to prevent such thermal runaway. In cases where adding emitter ballast resistors is not an option, instead ballast resistors can be added in the base. With ballasting resistors, as shown in Figure 11.51, the input voltage is applied across the base-emitter junction and the series resistors, so as V_{BE} and current increase, there is a larger voltage across the resistor, limiting the increase of V_{BE}.

It is also possible to decrease input voltage as temperature increases, for example, by using a diode in the input circuit, using a current mirror, or using a more complex arrangement of thermal sensors and bandgap biasing circuits. An example of a thermal biasing circuit for a push-pull class B output stage is shown in Figure 11.52. In this example, all diodes and transistors are assumed to be at the same temperature. As temperature rises, V_D falls, reducing V_{BE}, and keeping I constant.

For CMOS transistors, the required V_{GS} to keep a constant current as temperature increases is similar to the bipolar transistor at low V_{GS} and low currents with numbers as high as 2.4 mV/°C [13]. However, as a power amplifier operating at high V_{GS} and high current, the required V_{GS} increases for increasing temperatures, and thus CMOS does not suffer from the same thermal runaway problems.

However, it is important that the amplifier operates properly over temperature and as a result temperature compensation is required. With a current mirror style of biasing, if both the reference transistor and the driver transistor are at the same temperature, the current would remain roughly constant. Typically, more elaborate biasing schemes are used, incorporating bandgap references and current that is either constant with temperature or increasing with temperature to compensate for reduced performance at higher temperature. In some cases feedback is used to boost the bias voltage under high power to increase the performance of the amplifier [14].

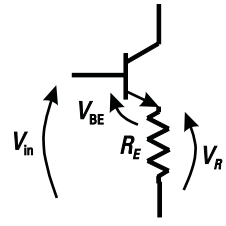

Figure 11.51 Balasting.

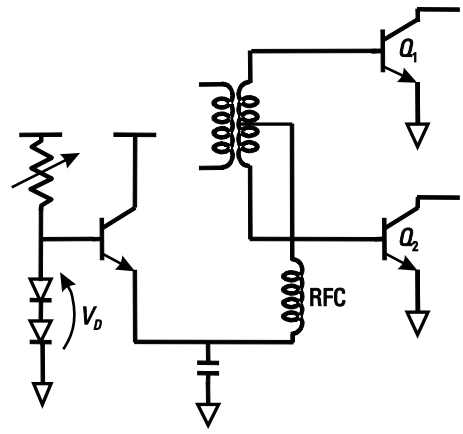

Figure 11.52 Temperature compensation.

11.18 Breakdown Voltage and Biasing

Avalanche breakdown occurs when the electric field within the depletion layer provides sufficient energy for free carriers to knock off additional valence electrons from the lattice atoms. These secondary electrons in turn generate more free carriers, resulting in avalanche multiplication. A measure of breakdown is $V_{CEO,\max}$, which is the maximum allowable value of V_{CE} with the base open circuited. A typical value in a 3-V process might be 5V. However, under ac conditions and with the base matched, swings of V_{CE} past $2V_{CC}$ can typically be provided safely [15].

Similar limitations occur in CMOS transistors. Possibly the most important is gate oxide breakdown due to high V_{GS} or V_{DS}. High electric fields result in tunneling current and ultimately a short circuit or low impedance path across the gate. Another effect is that of hot carriers in the channel occurring when there is simultaneously high V_{DS} and high I_{DS}. This can cause surface defects in the channel or charges trapped in the gate resulting in reduced performance. Another effect is *punch through,* which occurs when depletion regions around source and drain expand enough to meet or to be close together. This reduces the barrier to current flow, and current flows even without gate bias. On the drain characteristics showing I_{DS} versus V_{DS}, this shows up as an upward bend in the curves for high values of V_{DS}.

One of the ways of preventing breakdown is to limit the supply voltage and use several transistors in series, for example, cascode devices, to reduce the voltage across each transistor. Finally, it is also possible to make use of more complex biasing, which is adaptive or at least variable. As one example, in [14], a simple RC network connects the output back to the cascode gate and provides a bias voltage which tracks the output. In this way the instantaneous voltage across the cascode transistor is reduced, reducing the likelihood of breakdown. At the same time the variable bias allows the amplifier to be more linear and to have a higher PAE across a broader range of output power levels compared to an amplifier with fixed bias.

A complication during the design phase is that some transistor models are too simplistic and do not include the effects of breakdown. Clearly, while better simulation models are required, many designers depend on laboratory verifications, common sense, and experience.

11.19 Packaging

How does one remove heat from a power amplifier? One possible mechanism is thermal conduction through direct contact, for example, when the die is mounted directly on a metal backing. In flip-chip implementation, thermal conduction is through the solder bumps to the printed circuit board. It should be noted that wires to the package or directly to the printed circuit board (called *chip-on-board*) will remove a very limited amount of heat.

11.20 Effects and Implications of Nonlinearity

Linearity of the PA is important with certain modulation schemes. For example, filtered quadrature phase shift keying (QPSK) is often used largely because it can have a very narrow bandwidth. However, power amplifiers are required to be sufficiently linear to avoid spectral regrowth, which will dump power into adjacent bands. Note that with offset QPSK (OQPSK) and $\pi/4$-QPSK, the drawback is less severe because of the smaller phase steps. Minimum shift keying (MSK) modulation is typically a constant envelope and so allows the use of nonlinear power amplifiers; however, MSK requires wider bandwidth channels. FM and frequency shift keying (FSK) are two other constant envelope modulation schemes that can make use of high-efficiency power amplifiers.

In addition to spectral regrowth, nonlinearity in a dynamic system may lead to AM-PM conversion corrupting the phase of the carrier. Linearity is often checked using a two-tone test as previously described. However, this may not be realistic in predicting behavior when a real signal is applied. In such cases, it is possible to apply a modulated waveform and to measure the spectral regrowth.

11.20.1 Cross Modulation

Nonlinear power amplifiers can cause signals to be spread into adjacent channels, which can cause cross modulation. This is based on the same phenomena as third-order intermodulation for nonlinear amplifiers with two-tone inputs.

11.20.2 AM-to-PM Conversion

The phase response of an amplifier can change rapidly for signal amplitudes that result in gain compression, as illustrated in Figure 11.53. Thus, any amplitude variation (AM) in this region will result in phase variations (PM); hence, we can say there has been AM-to-PM conversion.

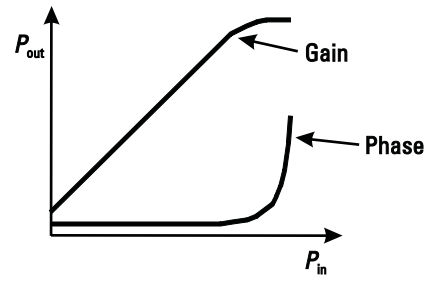

Figure 11.53 AM-to-PM conversion.

11.20.3 Spectral Regrowth

There can be additional problems that are worse for systems with varying envelopes. As an example, envelope variations can occur for modulation schemes that have zero crossings, such as binary phase shift keying (BPSK) or QPSK. [We note that /4 differential quadrature phase shift keying (DQPSK) or Gaussian minimum shift keying (GMSK) do not have zero crossings.] Due to band limiting, these zero crossings get converted into envelope variations. Any amplifier nonlinearities will cause spreading of frequencies into the adjacent channels, referred to as *spectral regrowth*. This is illustrated in Figure 11.54 for a QPSK signal with symbol period T_s.

11.20.4 Linearization Techniques

In applications requiring a linear PA (such as filtered QPSK or QAM, or systems carrying many channels such as base station transmitters or cable television transmitters), one can use a class A power amplifier at 30% to 40% efficiency, or a higher efficiency power amplifier operating in a nonlinear manner, but then apply linearization techniques. The overall efficiency reduction can be minimal while still reducing the distortion.

Linearization techniques tend to be used in expensive complex RF and microwave systems and less in low-cost portable devices, often because of the inherent complexities, the need to adjust, and the problems with variability of device characteristic with operating conditions and temperature. However, some recent papers, such as [14, 16], have demonstrated a growing interest in techniques to achieve enhanced linearity for integrated applications.

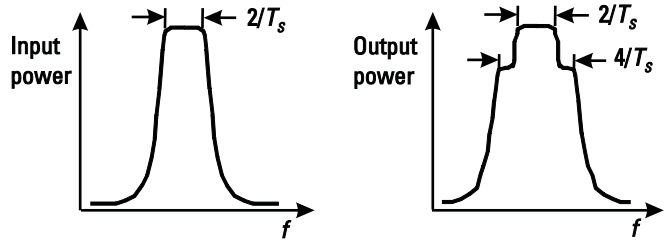

Figure 11.54 Spectral regrowth for QPSK signal with symbol period T_s.

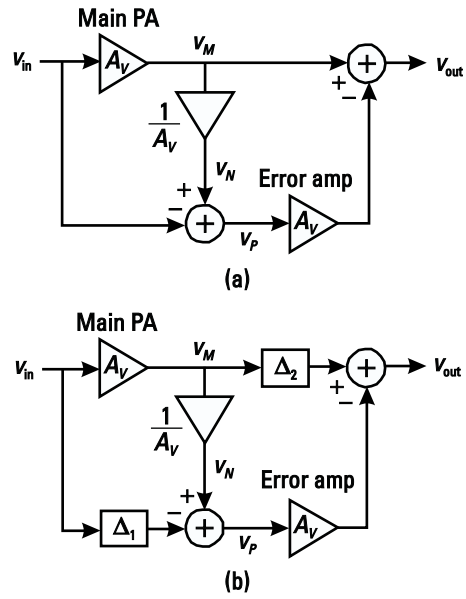

Figure 11.55 Feedforward linearization: (a) simple feedforward topology and (b) addition of delay elements.

11.20.5 Feedforward

The feedforward technique is shown in Figure 11.55. The amplifier output is $v_M = A_V v_{\text{in}} + v_D$, which consists of $A_V v_{\text{in}}$, the amplified input signal, and distortion components v_D, which we are trying to eliminate. This signal is attenuated to result in $v_N = v_{\text{in}} + v_D/A_V$. If this is compared to the original input signal, the result is $v_P = v_D/A_V$ and after amplification by A_V, the result is $v_Q = v_D$. If this is subtracted from the output signal, the result is $v_{\text{out}} = A_V v_{\text{in}}$ as desired.

At high power, there is significant phase shift, and thus phase shift has to be added as shown in Figure 11.55(b).

A major advantage of feedforward over feedback is that it is inherently stable in spite of finite bandwidth and high phase shift in each block. There are also a number of difficulties with the feedforward technique. The delay can be hard to implement, as it must be the correct value and should ideally have no loss. The output subtractor should also be low loss.

Another potential problem is that linearization depends on gain and phase matching. For example, if $A/A = 5\%$ and $\phi = 5°$, then intermodulation products are attenuated by only 20 dB.

11.20.6 Feedback

Successful feedback requires high enough gain to reduce the distortion, but low enough phase shift to ensure stability. These conditions are essentially impossible to obtain in a PA at high frequency. However, since a PA is typically an upconverted signal, if the output of the PA is first downconverted, the result can be compared to the original input signal. At low frequency, the gain and phase problems are less

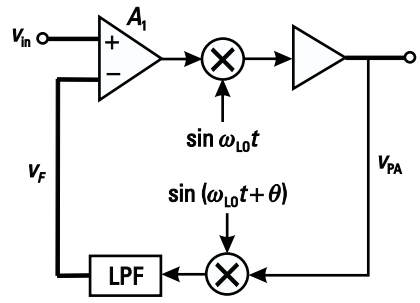

Figure 11.56 Feedback linearization techniques.

severe. An example of using such a feedback technique is shown in Figure 11.56. There have been a number of variations on this technique, including techniques to remove envelope variations. The interested reader is referred to [7].

11.20.7 Predistortion

Predistortion is now a common way to do linearization. The predistortion can be done open loop where the inverse of a known nonlinearity is applied. The predistortion can also be made adjustable or adaptive using feedback to adjust for temperature, process, or voltage variations. Digital signal processing techniques are often used allowing for maximum flexibility. Previously mentioned problems with feedback are avoided because the feedback is done very slowly so that instantaneous changes are not followed [17–19].

11.21 CMOS Power Amplifier Examples

Recently, a number of CMOS power amplifiers have been published [9–11, 15–22]. CMOS power amplifiers continue to be of interest for fully integrated single-chip radios, although a number of practical issues such as coupling and isolation are important for the commercial adoption of such radios. While CMOS may still not be the technology of choice for stand-alone power amplifiers, recently there have been several fully integrated designs demonstrating high output power. One of these [9] achieved 35 dBm of output power in a 130-nm CMOS process making use of input and output coupling and impedance matching using concentric coupled slab inductors and transformers with output transistors distributed along the length of the transformer. Another example, a class E power amplifier, achieved 30.5 dBm of output power using a narrowband lumped element balun, employing the minimum number of integrated inductors for minimum power loss [11]. Earlier, other CMOS power amplifiers were hybrid designs with a few components off chip, for example some of the input and output matching components. An example from 1997 shown in Figure 11.57 was a CMOS power amplifier in a 0.8-μm process at 900 MHz [23]. This was designed for a standard that had constant amplitude waveforms, so linearity was of less importance. The cascode input stage operates in class A (input is +5 dBm), the second stage operates in class AB, and the last two stages operate

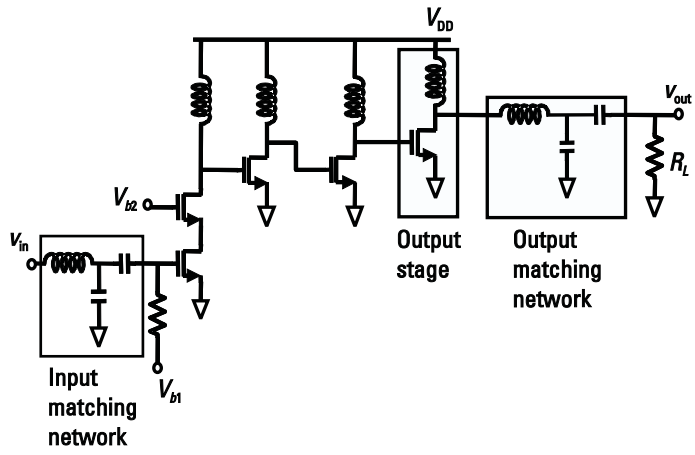

Figure 11.57 Power amplifier example.

as switching circuits to deliver substantial power with relatively high efficiency. (Note that class C amplifiers are also high efficiency, however, only at low conduction angles; thus, they provide high efficiency only at low power levels.) Measured results with a 2.5-V power supply showed an output power of 1W (30 dBm) with a power-added efficiency of 42%.

References

[1] Owler, T., et al., "Efficiency Improvement Techniques at Low Power Levels for Linear CDMA and WCDMA Power Amplifiers," *Proc. Radio Frequency Integrated Circuits Symposium*, Seattle, WA, May 2001, pp. 41–44.

[2] Staudinger, J., "An Overview of Efficiency Enhancements with Applications to Linear Handset Power Amplifiers," *Proc. Radio Frequency Integrated Circuits Symposium*, Seattle, WA, May 2001, pp. 45–48.

[3] Grebennikov, A. V., "Switched-Mode Tuned High-Efficiency Power Amplifiers: Historical Aspect and Future Prospect," *Proc. Radio Frequency Integrated Circuits Symposium*, Seattle, WA, May 2001, pp. 49–52.

[4] Krauss, H. L., C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, New York: John Wiley & Sons, 1980.

[5] Cripps, S. C., *RF Power Amplifiers for Wireless Communications*, 2nd ed., Norwood, MA: Artech House, 2006.

[6] Albulet, M., *RF Power Amplifiers*, Atlanta, GA: Noble Publishing, 2001.

[7] Kenington, P. B., *High Linearity RF Amplifier Design*, Norwood, MA: Artech House, 2000.

[8] Sokal, N. O., and A. D. Sokal, "Class E—A New Class of High Efficiency Tuned Single-Ended Power Amplifiers," *IEEE J. Solid State Circuits*, Vol. SC-10, No. 3, June 1975, pp. 168–176.

[9] Aoki, I., et al., "A Fully-Integrated Quad-Band GSM/GPRS CMOS Power Amplifier," *IEEE J. Solid State Circuits*, Vol. 43, No. 12, December 2008, pp. 2747–2758.

[10] Haldi, P., et. al., "A 5.8 GHz 1 V Linear Power Amplifier Using a Novel On-Chip Transformer Power Combiner in Standard 90 nm CMOS," *IEEE J. Solid-State Circuits*, Vol. 43, No. 5, May 2008, pp. 1054–1063.

[11] Brama, R., et al., "A 30.5 dBm 48% PAE CMOS Class-E PA with Integrated Balun for RF Applications," *IEEE J. Solid-State Circuits*, Vol. 43, No. 8, August 2008, pp. 1755–1762.

[12] Gonzalez, G., *Microwave Transistor Amplifiers, Analysis and Design,* 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 1997.

[13] Baker, R. J., *CMOS Circuit Design, Layout, and Simulation,* 2nd ed., New York: Wiley-IEEE Press, 2004.

[14] Sowlati, T., and D. Leenaerts, "A 2.4GHz, 0.18μm CMOS Self-Biased Cascode Power Amplifier with 23dBm Output Power," *Proc. ISSCC*, 2002, February 2002, pp. 294–295.

[15] Veenstra, H. G., et al., "Analysis and Design of Bias Circuits Tolerating Output Voltage Above BV_{CEO}," *IEEE J. Solid-State Circuits*, Vol. 40, No. 10, October 2005, pp. 2008–2018.

[16] Shinjo, S., et al., "A 20mA Quiescent Current CV/CC Parallel Operation HBT Power Amplifier for W-CDMA Terminals," *Proc. Radio Frequency Integrated Circuits Symposium*, Seattle, May 2001, pp. 249–252.

[17] Kusunoki, S., et al., "Power-Amplifier Module with Digital Adaptive Predistortion for Cellular Phones," *Trans. Microwave Theory and Techniques*, Vol. 50, No. 12, December 2002, pp. 2979–2986.

[18] Kidwai, A. A., et al., "Fully Integrated 23dBm Transmit Chain with On-Chip Power Amplifier and Balun for 802.11a Application in Standard 45nm CMOS Process," *Proc. Radio Frequency Integrated Circuits Symposium*, Boston, MA, May 2009, pp. 273–276.

[19] Presti, C. D., et al., "A 25 dBm Digitally Modulated CMOS Power Amplifier for WCDMA/EDGE/OFDM with Adaptive Digital Predistortion and Efficient Power Control," *IEEE J. Solid-State Circuits*, July 2009, pp. 1883–1896.

[20] Yoo, C., and Q. Huang, "A Common-Gate Switched 0.9W Class E Power Amplifier with 41% PAE in 0.25 m CMOS," *IEEE J. Solid-State Circuits*, May 2001, pp. 823–830.

[21] Kuo, T., and B. Lusignan, "A 1.5W Class-F RF Power Amplifier in 0.25 m CMOS Technology," *Proc. ISSCC*, 2001, February 2001, pp. 154–155.

[22] Su, D., et al., "A 5GHz CMOS Transceiver for IEEE 802.11a Wireless LAN," *Proc. ISSCC*, 2002, February 2002, pp. 92–93.

[23] Su, D., and W. McFarland, "A 2.5-V, 1-W Monolithic CMOS RF Power Amplifier," *Proc. Custom Integrated Circuits Conference,* May 1997, pp. 189–192.

About the Authors

John W. M. Rogers received a Ph.D. in electrical engineering from Carleton University, Ottawa, Canada, in 2002. Concurrent with his Ph.D. research, he worked as part of a design team with SiGe Semiconductor that developed a cable modem IC for the DOCSIS standard. From 2002 to 2004, he collaborated with Cognio Canada Ltd. doing research on MIMO RFICs for WLAN Applications. Since 2002 he has been a member of the faculty of engineering at Carleton University, where he is now an associate professor. From 2007 to 2008 he was on sabbatical and working with Alereon Inc. developing ICs for UWB applications. He is the coauthor of *Radio Frequency Integrated Circuit Design* (Artech House, 2003) and *Integrated Circuit Design for High Speed Frequency Synthesis* (Artech House, 2006). His research interests are in the areas of RFIC and mixed-signal design for wireless applications. Dr. Rogers was the recipient of an IBM faculty partnership award in 2004, and an IEEE Solid-State Circuits Predoctoral Fellowship in 2002. He holds five U.S. patents and is a member of the Professional Engineers of Ontario and is a senior member of the IEEE. He has been serving as a member of the technical program committee for the Custom Integrated Circuits Conference since 2006 and the BiCMOS Circuits and Technology Meeting since 2008.

Calvin Plett received a B.A.Sc. in electrical engineering from the University of Waterloo, Canada, in 1982, and an M.Eng. and a Ph.D. from Carleton University, Ottawa, Canada, in 1986 and 1991, respectively. Prior to 1982, he worked for a number of companies including Atomic Energy of Canada, Xerox, Valcom, Central Dynamics, and Philips. From 1982 to 1984 he worked with Bell-Northern Research doing analog circuit design. In 1989 he joined the Department of Electronics at Carleton University, Ottawa, Canada, where he is now a professor. He has done consulting work for Nortel Networks in RFIC design and has been involved in collaborative research that involved numerous graduate and undergraduate students and various companies including Nortel Networks, SiGe Semiconductor, Philsar, Conexant, Skyworks, IBM, and Gennum. He has authored or coauthored over 85 technical papers that have appeared in international journals and conferences. He is a coauthor of the books *Radio Frequency Integrated Circuit Design* (Artech House, 2003) and *Integrated Circuit Design for High-Speed Frequency Synthesis* (Artech House, 2006). His research interests include the design of analog and radio-frequency integrated circuits, including filter design, and communications applications. Dr. Plett is a member of AES and the PEO and a senior member of the IEEE. He has been the faculty advisor to the student branch of the IEEE at Carleton University for about 18 years. He coauthored papers that won the best student paper awards at BCTM 1999 and at RFIC 2002.

Index

N

Q